electrónica

MDPI

Artículo

Predicción de las características de los enlaces serie de alta velocidad Basado en una red neuronal profunda (DNN): Transformer Modelo en cascada

Liyin Wu [1], Jingyang Zhou [1], Haining Jiang [1], Xi Yang [1], Yongzheng Zhan [2] y Yinhang Zhang [1,*]

[1] Escuela de Ingeniería Electrónica y Comunicaciones, Universidad de Jishou, Jishou 416000, China; 2022700814@stu.jsu.edu.cn (LW); 2022700812@stu.jsu.edu.cn (JZ); 2023700812@stu.jsu.edu.cn (HJ); ynkej@163.com (XY)

[2] Shandong Yunhai Guochuang Cloud Computing Equipment Industry Innovation Co., Ltd., Jinan 250101, China; sdzyz1989@163.com *

Correspondencia: zyh149032@jsu.edu.cn

Resumen: El nivel de diseño de las características físicas del canal tiene una influencia crucial en la calidad de transmisión de enlaces seriales de alta velocidad. Sin embargo, el diseño de canales requiere un complejo proceso de simulación y verificación. En este artículo, se propone un modelo de red neuronal en cascada construido a partir de una red neuronal profunda (DNN) y un transformador. Este modelo toma características físicas como entradas e importa una respuesta de un solo bit (SBR) como conexión, que se mejora mediante la predicción de características de frecuencia y parámetros del ecualizador. Al mismo tiempo, el análisis de la integridad de la señal (SI) y la optimización del enlace se logran prediciendo diagramas de ojo y márgenes operativos del canal (COM). Además, la optimización bayesiana basada en el proceso gaussiano (GP) se emplea para la optimización de hiperparámetros (HPO). Los resultados muestran que el modelo en cascada DNN-Transformer logra predicciones de alta precisión de múltiples métricas en la predicción y optimización del rendimiento, y el error relativo máximo de los resultados del conjunto de pruebas es inferior al 2% bajo la arquitectura de ecualizador de un TX de 3 grifos. FFE, un RX CTLE con doble ganancia de CC y un RX DFE de 12 toques, que es más potente que otros modelos de aprendizaje profundo en términos de capacidad de predicción.

Palabras clave: enlace de alta velocidad; integridad de la señal; diagrama de ojo; margen operativo del canal; modelo en cascada

Cita: Wu, L.; Zhou, J.; Jiang, H.; Yang, X.; Zhan, Y.; Zhang, Y. Predecir las características de los enlaces serie de alta velocidad basados en una red neuronal profunda (DNN)— Modelo de transformador en cascada. Electrónica 2024, 13 , 3064. https://doi.org/10.3390/electronics13153064

Editor académico: Shing-Hong Liu

Recibido: 1 de julio de 2024
Revisado: 24 de julio de 2024
Aceptado: 31 de julio de 2024
Publicado: 2 de agosto de 2024

1. Introducción

A medida que el ancho de banda de transmisión de la tecnología de enlace serie alámbrico alcanza el nivel de GHz, ya no es posible garantizar una transmisión de señal eficiente simplemente optimizando el dieléctrico y la estructura de diseño. Los sistemas de enlace serie de alta velocidad sufren graves problemas de integridad de la señal (SI) debido al efecto superficial, la pérdida dieléctrica, la diafonía, las reflexiones y la fluctuación; por lo tanto, el análisis SI se vuelve cada vez más estricto en la etapa de diseño de enlaces seriales de alta velocidad. El análisis de simulación de SI generalmente consta de dos pasos: solucionadores de campos electromagnéticos (EMFS) y simulación de sistemas de circuitos [1]. En primer lugar, se utiliza un EMFS para obtener parámetros S para caracterizar la respuesta de frecuencia del circuito. Luego, estos parámetros S se importan al sistema de circuito modelo para la simulación en el dominio del tiempo para obtener las principales métricas del SI, incluido un diagrama de ojo, la respuesta al impulso y las formas de onda transit

Aunque el análisis SI tradicional basado en modelos físicos de enlaces de alta velocidad puede ofrecer una gran precisión, consume mucho tiempo y recursos informáticos. La interfaz de modelo algorítmico de especificación de información del búfer de entrada/salida (IBIS-AMI) es un modelo de comportamiento que simula el comportamiento y los algoritmos de entrada/salida de enlaces serie de alta velocidad de extremo a extremo , simplificando los detalles físicos internos [2]. En comparación con los modelos físicos, tiene las ventajas de velocidad, simplicidad y bajo consumo de recursos, pero tiene poca precisión y carece de flexibilidad. La comparación de plantillas de estándares de cumplimiento es otro método comúnmente utilizado [3]. Esto puede estimar rápida e intuitivamente el rendimiento del canal, pero inevitab

descarta los márgenes del canal debido a la necesidad de satisfacer estrictamente métricas discretas [4]. Por el contrario, la técnica del margen operativo del canal (COM) puede superar eficazmente esta desventaja buscando el espacio de diseño óptimo de todo el enlace en forma de relación señal-ruido (SNR). En comparación con las métricas SI tradicionales, como los diagramas de ojo y la tasa de error de bits (BER), el enfoque COM muestra ventajas como una operación más simple, una velocidad más rápida y pruebas más eficientes. Obviamente, en comparación con la evaluación de canales del Anexo69B para Ethernet de 10 Gb/s (10 GbE), el uso de COM es más preciso [5]. Aunque el enfoque COM puede proporcionar una forma eficaz de evaluar y optimizar enlaces de alta velocidad, requiere numerosas iteraciones para la búsqueda espacial y no es lo suficientemente flexible para analizar canales de alta capacidad.

El aprendizaje automático (ML) se ha utilizado ampliamente en la simulación y el diseño de enlaces de alta velocidad en los últimos años y muestra la capacidad de mejorar la eficiencia del análisis SI. Los autores de [4,6–10] buscaron características a partir de datos de simulación y entrenaron modelos de red neuronal artificial (ANN), red neuronal profunda (DNN) y máquina de vectores de soporte de mínimos cuadrados (LS-SVM) para reemplazar las simulaciones de sistemas de circuitos. para la predicción precisa de métricas en el dominio del tiempo (TD) y la frecuencia, como la altura del ojo (EH)/ ancho del ojo (EW) y la pérdida de retorno (RL)/pérdida de inserción (IL); sin embargo, la adquisición de datos de simulación consume muchos recursos computacionales y tiempo. Las redes neuronales de avance (FNN), la regresión de bosque aleatoria (RFR) y las máquinas de vectores de soporte (SVM) se utilizan para lograr predicciones simples de datos de simulación [11-13], como para predecir parámetros S y respuestas de impulso. Cuando se trata de datos de simulación más complejos, los métodos tradicionales de aprendizaje automático pueden perder muchas funciones. La red neuronal recurrente (RNN) es un modelo de ML que puede capturar de manera efectiva las características de datos complejos, especialmente para datos de secuencia como parámetros S y respuestas de impulso. Los autores de [14–20] emplean la arquitectura RNN y la memoria larga a corto plazo (LSTM) para crear modelos sustitutos para predecir las formas de onda de respuesta transitoria de enlaces complejos de alta velocidad. Estos modelos sustitutos independientes pueden realizar análisis SI basados en datos de simulación, pero no pueden abordar los parámetros físicos de los enlaces. El análisis SI basado en parámetros físicos requiere métodos de aprendizaje automático más complicados, y los autores de [21-24] lograron predicciones efectivas de los parámetros físicos utilizados para evaluar el rendimiento del enlace de alta velocidad combinando múltiples algoritmos de aprendizaje profundo. Además, se puede lograr una optimización de la arquitectura equilibrada para enlaces serie de alta velocidad basándose en ML [4,25-27]. En conclusión, la predicción del rendimiento y la optimización de la arquitectura [28] son dos partes importantes de las aplicaciones de ML para e

En este artículo, se propone un modelo en cascada DNN-Transformer para el análisis SI y la optimización de enlaces seriales de alta velocidad. Este modelo puede omitir el uso de EMFS y simulación de sistemas de circuitos, y predice directamente métricas SI, incluidas EH/EW, IL/RL, respuesta de impulso y valores COM, según parámetros físicos. Mientras tanto, la optimización de los enlaces se puede lograr utilizando este modelo para predecir los valores COM correspondientes y los parámetros del ecualizador para los enlaces con diferentes configuraciones de ecualizador. Además , la optimización bayesiana basada en el proceso gaussiano (GP) se utiliza para optimizar los hiperparámetros en las mismas condiciones para diferentes combinaciones de modelos. En comparación con la predicción del rendimiento del enlace de alta velocidad utilizando ML tradicional [6-10], DNN-Transformer puede utilizar directamente los parámetros físicos del enlace para el análisis y su precisión de predicción es significativamente mejor. Para la predicción de datos de simulación, como respuestas a impulsos, DNN-Transformer puede capturar más características y su capacidad de predicción es más precisa que la que se logra cuando se utiliza un modelo RNN [14] o LSTM [20] solo para crear un modelo sustituto. Los resultados muestran que el modelo DNN-Transformer puede lograr una predicción más efectiva. El rendimiento y la viabilidad del método propuesto se ilustran mediante los datos de predicción y los resultados gráficos que se presentan en este documento.

Las principales contribuciones de este artículo se pueden resumir de la siguiente

manera: (1) Con base en los parámetros físicos clave de los canales, se utilizan modelos de redes neuronales para analizar directamente el rendimiento del enlace, y esto evita los procesos que requieren mucho tiempo de resolución de EMF y simulación de sistemas de circuitos. .

Las principales contribuciones de este artículo se pueden resumir de la siguiente manera:

(1) Con base en los parámetros físicos clave de los canales, se utilizan modelos de redes neuronales para analizar directamente el rendimiento del enlace, y esto evita los procesos que requieren mucho tiempo de resolución de EMF y simulación de sistemas de circuitos.

(2) A través de modelos de redes neuronales, la predicción precisa de múltiples indicadores SI y del ecualizador de enlace. Además, mostramos que el análisis SI y el enlace La optimización también se puede lograr rápidamente.

(3) Se propone un modelo en cascada DNN-Transformer, se utiliza la optimización bayesiana para sintonice los hiperparámetros de este modelo y se demuestra su rendimiento superior. comparándolo con otros modelos.

El resto de este documento está organizado de la siguiente manera: La Sección 2 analiza los principios básicos funciona este trabajo, incluidos los principios de los métodos SI y COM. La sección 3 describe cómo Transformer, el conjunto de datos de simulación. En la Sección 4, la idea de diseño del DNN-Transformer el modelo y se presentan las métricas de prueba. La sección 5 proporciona los resultados numéricos para demostrar predicción de nuestro modelo. Finalmente, la Sección 6 concluye el artículo.

2. Métodos Fundamentales

2.1. Integridad de la señal y parámetros S

A medida que aumentan las velocidades de transmisión, las ineficiencias de los métodos tradicionales de transmisión de señales en paralelo se están volviendo evidentes y los métodos de transmisión digital en serie de alta velocidad siendo ampliamente utilizada. La Figura 1 muestra la estructura SerDes ampliamente utilizada en una estructura de alta velocidad. Sistema de enlace en serie. Consta de un chip transmisor con un serializador y un ecualizador feedforward (FFE), un canal y un chip receptor con un ecualizador de retroalimentación de decisión (DFE) y un deserializador. El FFE en el chip transmisor se utiliza principalmente para realzar la señal para mejorar su componente de alta frecuencia. La composición de un canal alámbrico es compleja, normalmente incluye pistas, vías y conectores, etc. Estas estructuras físicas exhiben un filtro de paso bajo y pueden aumentar la pérdida de alta frecuencia de una señal, empeorando gravemente la calidad de transmisión de la señal y provocar problemas de SI. El ecualizador lineal de tiempo continuo (CTLE) en el extremo receptor se utiliza principalmente para compensar la pérdida de alta frecuencia del canal, eliminar los precursores y suprimir el seguimiento de la respuesta del pulso. El DFE se utiliza principalmente para eliminar los poscursores de la respuesta del pulso. La combinación y la configuración de parámetros del FFE, CTLE y DFE tienen un impacto significativo en el SI de los enlaces de comunicación serie por cable de ultra alta velocidad.

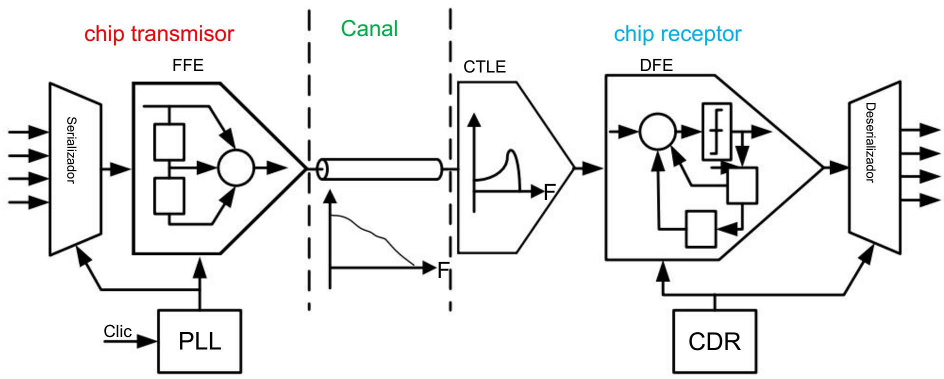
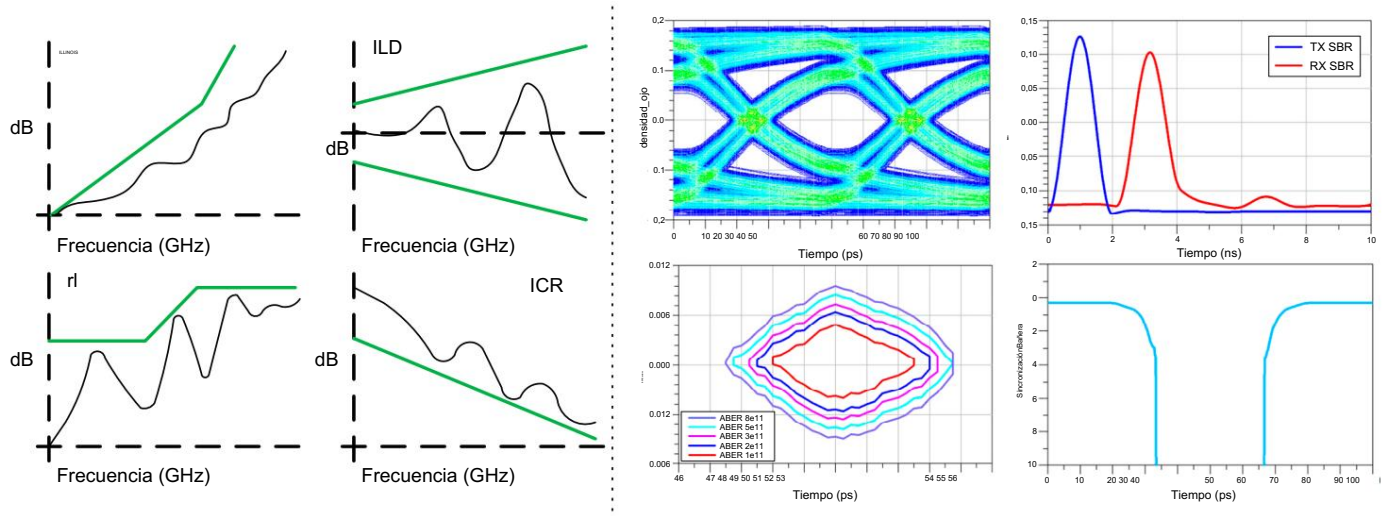


Figura 1. Estructura común de un circuito SerDes.

El objetivo principal del análisis SI es detectar y reducir los factores que causan pérdidas, tales como la reflexión y la diafonía. Por un lado, el SI puede analizarse desde la frecuencia, como se muestra en la Figura 2a. Esto incluye principalmente IL, RL, pérdida de inserción desviación de la pérdida de inserción (ILD) y la relación pérdida de inserción/diafonía (ICR), que posteriormente se colocan en una plantilla para su evaluación. Por otro lado, el SI se puede evaluar en el dominio del tiempo, como a través de diagramas de ojo, simulaciones transitorias, curvas de bañera y tasas de error, como se muestra en la Figura 2b.

tasas de error de bits, como se muestra en la Figura 2b.

El método crítico del análisis SI tradicional implica la adquisición de parámetros S. Estos parámetros contienen características integrales del dominio de frecuencia (FD) del canal de transmisión y ofrecen una gran cantidad de información sobre aspectos como la reflexión ción, diafonía y pérdida. Además, los parámetros S se pueden emplear en simulaciones en el dominio del tiempo para generar datos como diagramas de ojo y curvas de bañera.



(a) (b)

Figura 2. (a) FD. (b) TD. Ilustración de dos métodos de simulación diferentes para detectar SI.

Figura 2. (a) FD. (b) TD. Ilustración de dos métodos de simulación diferentes para detectar SI.

2.2. Margen operativo del canal ES

El método crítico del análisis SI tradicional implica la adquisición de parámetros S. Estos parámetros contienen características integrales del dominio de frecuencia (FD) del canal de transmisión y que ofrece una gran cantidad de información sobre aspectos como la reflexión, diafonía y pérdida. Además, los parámetros S se pueden emplear en simulaciones en el dominio del tiempo. para generar datos como diagramas de ojo y curvas de bañera.

El método COM es un método de caracterización de enlace serie de alta velocidad recomendado por el grupo de trabajo IEEE 802.3 para pruebas de cumplimiento de canales. La definición oficial es establece que un "COM es una figura de mérito (FOM) para canales determinados a partir de un mínimo arquitectura PHY de referencia y parámetros s del canal" [29]. Este enfoque proporciona un entorno relativamente preciso y justo para el diseño físico de canales considerando varios factores como pérdida, reflexión, interferencia entre símbolos (ISI), dispersión ISI. El método COM diafonía y especificaciones del dispositivo, lo que permite una evaluación relativamente precisa del canal, rendimiento y el impacto de los canales del agresor en los canales de la víctima. El impacto de que un Los ecualizadores se han considerado en versiones posteriores del método COM, y el valor de la arquitectura PHY de un COM basado en el FOM se puede mejorar seleccionando la configuración de los parámetros del En consecuencia, calcular el COM también puede determinar si la calidad del canal cumple con factores los requisitos SI del transceptor [30]. Como se muestra en la Ecuación (1), el COM puede ser ex- presionado por la relación entre la amplitud de la señal disponible y la amplitud del ruido estadístico An .

El método COM es un método de caracterización de enlace serie de alta velocidad recomendado por el grupo de trabajo IEEE 802.3 para pruebas de cumplimiento de canales. La definición oficial establece "COM sea una figura de mérito (FOM) para canales determinada a partir de una referencia mínima y los parámetros s del canal" [29]. Este enfoque proporciona una relativamente ecualizador, entorno preciso y justo para el diseño físico de canales considerando varios como pérdida, reflexión, interferencia entre símbolos (ISI), dispersión ISI, diafonía y especificaciones del dispositivo, lo que permite una evaluación relativamente precisa del rendimiento del canal y el impacto de los canales del agresor en los canales de la víctima. El impacto de los empates ha sido considerado en versiones posteriores del método COM, y el valor de un COM basado en el FOM se puede mejorar seleccionando la configuración de los parámetros del ecualizador. Como consecuencia, calcular el COM también puede determinar si la calidad del canal cumple con los requisitos del transceptor. Requisitos SI [30]. Como se muestra en la Ecuación (1), el COM se puede expresar mediante la relación de

$$COM = 20 \text{ registro}_{10} \left(\frac{A_s}{An} \right) \quad (1)$$

El proceso de derivar el COM implica varios pasos esenciales, incluida la determinación de la extraer la función de transferencia, convertir la función de transferencia en respuesta al impulso, aplicar algoritmos de ecualizador del transmisor y receptor y realizar cálculos estadísticos de ruido. La Figura 3 presenta un diagrama de flujo de la derivación del modelo COM y muestra dos Como COM = 20 × log10 (1) Vías para la víctima y canales para el agresor. Además, el modelo también considera la interferencia del paquete del transmisor y del receptor S (tp/rp), la resistencia de terminación Rd, ecualizadores, El proceso Un de derivar el COM implica varios pasos esenciales, incluidos disuasorios y filtros. Para simular mejor los canales reales, el modelo COM incorpora minería gaussiana de la función de transferencia, convirtiendo la función de transferencia en respuesta al impulso, ruido blanco y fluctuación en el extremo del receptor. La Figura 4 muestra el proceso detallado de COM aplicando algoritmos de ecualizador de transmisor y receptor y realizando estadísticas de ruido. cálculo del valor. cálculos. La Figura 3 presenta un diagrama de flujo de la derivación del modelo COM y muestra dos Vías para la víctima y canales para el agresor. Además, el modelo también considera la interferencia del paquete del transmisor y del receptor S (tp/rp), la resistencia de terminación Rd, ecualizadores, y filtros. Para simular mejor los canales reales, el modelo COM incorpora Gaussian ruido blanco y fluctuaciones en el extremo del receptor. La Figura 4 muestra el proceso detallado de COM. cálculo del valor.

representa la variación del ruido en el transmisor. σISI denota la variación en la amplitud residual de ISI y σj indica la variación en la amplitud de la fluctuación de fase. En el análisis COM, la amplitud ISI de la fluctuación de fase se tiene en cuenta convirtiendo la fluctuación de fase horizontal en ruido vertical en el instante de muestreo. Además, σXTK representa la variación total de la diafonía de todas las rutas de interferencia, mientras que σN denota el ruido blanco gaussiano en el punto de muestreo del receptor.

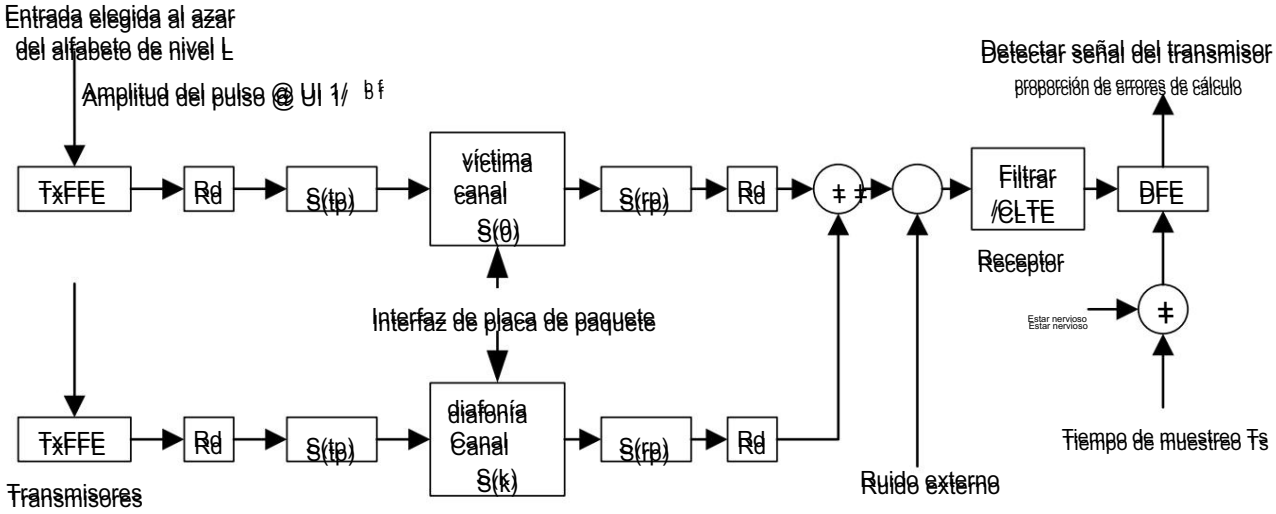


Figura 3. Un modelo COM típico.

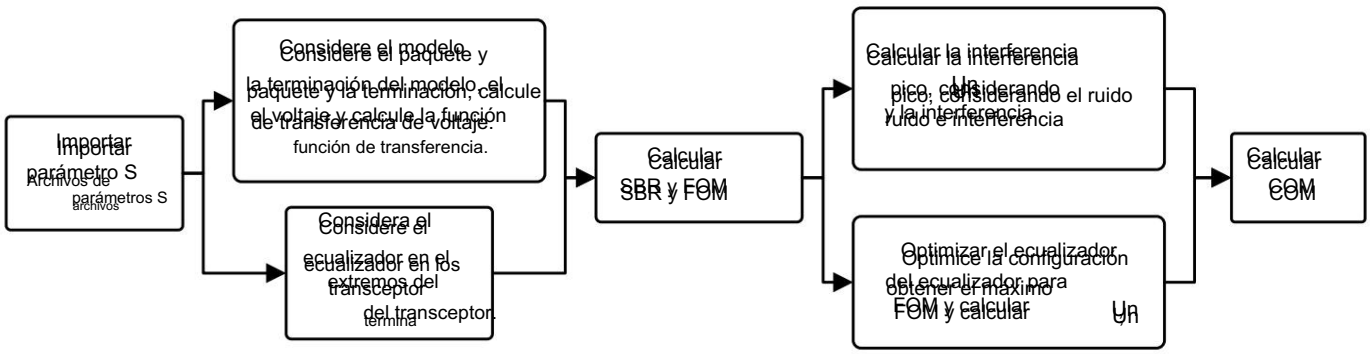


Figura 4. Proceso de cálculo de COM.

El enfoque COM implica explorar todas las combinaciones posibles de parámetros del TX y ecualizadores RX dentro de un rango establecido para encontrar la configuración que maximice el FOM. Este proceso determina los valores de As y An que dan como resultado el FOM óptimo.

Los métodos tradicionales se basan en varios indicadores, como la inquietud, la altura de los ojos y la fórmula de cálculo de los ojos, que pueden expresarse de varias maneras, como la fluctuación, la altura y el ancho del ojo, pero el COM sirve como una métrica integral para evaluar un enlace en serie. Puede ampliarse, pero el COM sirve como una métrica integral para evaluar un enlace en serie. Puede

$$FOM = 10 \times \log 10 \quad \frac{A_s^2}{\sigma_{Texas}^2 + \sigma_{ISI}^2 + \sigma_j^2 + \sigma_{XTK}^2 + \sigma_{norte}^2} \tag{2}$$

El numerador As se deriva de la amplitud de la respuesta al impulso en ts , que corresponde al cursor principal de la respuesta al impulso h(t). El denominador incluye la suma de las variaciones de todos los componentes de ruido, fluctuación e interferencia. σ_{TX}^2 representa el variación de ruido en el transmisor, σ_{ISI}^2 denota la varianza en la amplitud residual de ISI, y σ_j^2 indica la variación en la amplitud de la fluctuación. En el análisis COM, se tiene en cuenta la fluctuación convirtiendo la fluctuación horizontal en ruido vertical en el instante de muestreo ts . Además, σ_{XTK}^2 representa la variación total de la diafonía de todas las rutas de interferencia, mientras que σ_{norte}^2 denota el ruido blanco gaussiano en el punto de muestreo del receptor.

El enfoque COM implica explorar todas las combinaciones posibles de parámetros del TX y ecualizadores RX dentro de un rango establecido para encontrar la configuración que maximice el FOM. Este proceso determina los valores de As y An que dan como resultado el FOM óptimo.

Los métodos tradicionales se basan en varios indicadores como la inquietud, la altura de los ojos y la altura de los ojos. ancho, pero el COM sirve como una métrica integral para evaluar un enlace serie. Él puede reducir significativamente el tiempo de cálculo y el número de iteraciones, y proporciona una Evaluación precisa del rendimiento del canal antes y después de la ecualización.

reduce significativamente el tiempo de cálculo y el número de iteraciones, y proporciona una evaluación precisa del rendimiento del canal antes y después de la ecualización.

3. Construcción del conjunto de datos

3. Construcción del conjunto de datos

3.1. Diseño de canales y división de conjuntos de datos

3.1. Diseño de canales y división de conjuntos de datos

En este trabajo, los conjuntos de datos del canal se generaron utilizando un software de diseño de PCB profesional.

En este trabajo, los conjuntos de datos del canal se generaron utilizando un diseño de PCB profesional
mercancía. Como se muestra en la Figura 5, software Altium Designer 20.0.13 y Allegro 17.4
(software EDA). Como se muestra en la Figura 5, Altium Designer 20.0.13 y Allegro 17.4 (software EDA)
se utilizaron principalmente para crear los diseños de líneas diferenciales de alta velocidad. Los subse- se utilizaron
principalmente para crear diseños para líneas diferenciales de alta velocidad. Después,
Posteriormente, los archivos de diseño de PCB se importaron al Advanced Design System 2021 (ADS) o los archivos de
diseño de PCB se importaron al Advanced Design System 2021 (ADS) o HFSS 2023
HFSS 2023 R1 para simulación electromagnética (EM). Nuestro diseño se ajusta a la especificación R1 para simulación
electromagnética (EM). Nuestro diseño se ajusta a las especificaciones para USB
clones para líneas de transmisión diferencial de extremo a extremo USB, incluidos parámetros como líneas de transmisión
diferencial de extremo a extremo, incluidos parámetros como ancho de línea y
ancho y largo, la impedancia característica de los canales de transmisión y longitud diferencial, la impedancia característica
de los canales de transmisión y el espaciado diferencial de líneas.
Espaciado entre líneas.

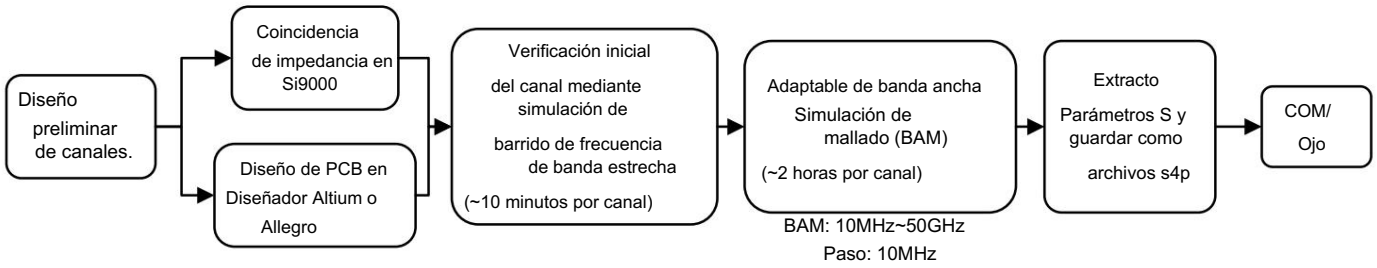


Figura 5. Proceso de simulación y fabricación de canales.
Figura 5. Proceso de simulación y fabricación de canales.

Las características diferenciadas de los datos del canal, como se presentan en la Tabla 1, principalmente
Las características diferenciadas de los datos del canal, como se presentan en la Tabla 1, principalmente
incluya parámetros como material, dimensiones de traza y tipo de PCB. Las características de la placa incluyen parámetros
como el material, las dimensiones de la traza y el tipo de PCB. Las características de la placa abarcan ambas características
numéricas, como la permitividad (Er), las características de disipación abarcan ambas características numéricas, como la
permitividad (Er), el factor de disipación.
tan (Df) y espesor de la placa, y características categóricas como el tipo de PCB. Er consiste (Df) y el grosor de la placa, y
características categóricas como el tipo de PCB. Er consiste en
permitividad de la lámina y el ángulo tangente de pérdida dieléctrica (TanD). Df se deriva
de la parte real de la permitividad de la lámina y el ángulo tangente de pérdida dieléctrica (TanD). Df es la parte real de la
derivado del modelo Djordjevic, que se define mediante parámetros que incluyen el bajo y del modelo Djordjevic, que se define mediante parámetros que incluyen el bajo y el alto
permitividad relativa de alta frecuencia. En este estudio, Df se cuantifica
altas frecuencias de TanD y permitividad relativa de alta frecuencia. En este estudio, Df son las frecuencias de TanD y la
TanD y sus correspondientes altas y bajas frecuencias.
cuantificado según los valores numéricos de TanD y sus correspondientes altas y bajas según los valores numéricos de
representado por un valor numérico
frecuencias. El tipo de PCB se clasifica en el tipo stripline, representado por un El tipo de PCB se clasifica en el tipo stripline,
representado por 0. valor numérico de 1, o el tipo microstrip, representado por 0. de 1, o el tipo microstrip,

Estas características de la traza se cuantifican en función de parámetros específicos, como la longitud, el ancho, el espesor, la
conductividad, el perfil y el espaciado diferencial de las líneas de la traza. El perfil es la Tabla 1. Parámetros de características del
caracterizado por la forma geométrica de la traza, como el número de esquinas. La otra entrada salida
canal y métodos cuantitativos.
Las características están representadas directamente por sus correspondientes valores numéricos.

Tabla 1. Parámetros de características del canal y métodos cuantitativos.

Variable de característica		Producción
Eh	Permitividad de la hoja de PCB parte real y TanD	
df	Factor de disipación de la hoja de PCB	
Rastro	Longitud de la traza de PCB de entrada Ancho de traza de PCB Espacio de línea diferencial Longitud de la traza de PCB Espesor de la traza de PCB Ancho de traza de PCB Conductividad de traza de PCB Estructura de traza de PCB Espaciado de líneas diferenciales (perfil)	diagramas de ojos valores COM Parámetros del ecualizador
Rastro	Longitud de la traza de PCB de entrada Permitividad de la hoja de PCB parte real y TanD Factor de disipación de la hoja de PCB Espacio de línea diferencial Longitud de la traza de PCB Espesor de la traza de PCB Ancho de traza de PCB Conductividad de traza de PCB Estructura de traza de PCB Espaciado de líneas diferenciales (perfil)	diagramas de ojos valores COM Parámetros del ecualizador
zd	Tipo de PCB de impedancia	
Tipo de PCB	Stripline es 1 y microstrip es 0	

zd

diferencial Stripline es 1 y microstrip es 0. Estas características de traza se
cuantifican en función de parámetros específicos, como la traza.

longitud, ancho, espesor, conductividad, perfil y espaciado diferencial de líneas. El perfil es
caracterizado por la forma geométrica de la traza, como el número de esquinas. El otro

Las características están representadas directamente por sus correspondientes valores numéricos.

3.2. Configuración COM y ajuste del ecualizador

Este trabajo utiliza el programa versión COM 4.0 proporcionado por IEEE 802.3. La configuración hace referencia a la configuración de simulación COM mejorada (eCOM) con el USB4 Gen4 estándar e incorpora los elementos del estándar 50 GBASE-KR. Los ajustes de configuración de PAM4, como se muestran en la Tabla 2, emplean una combinación de un transmisor FFE y CTLE y DFE del lado del receptor. Esta configuración incluye los coeficientes de tap del TX FFE y RX DFE, el rango adaptativo de ganancia de CC y las posiciones de polo cero del RX CTLE. Los parámetros adicionales incluyen el número de niveles de señal, indicado por L. Para señales PAM4, L está configurado en 4. Según la configuración PAM4 recomendada basada en USB4, el símbolo La velocidad se establece en 20 GBd y la salida de voltaje pico diferencial del transmisor se establece en 0,4 V, denominado Av. El estándar USB4 Gen4 sugiere que se debe utilizar PAM3 como método de modulación. La configuración COM para PAM3 es aproximadamente la misma que la de PAM4, excepto que la velocidad de símbolo se establece en 25,6 GBd y L se establece en 3.

Tabla 2. Configuración de parámetros COM.

Parámetro	símbolo	Configuración	Unidad
Velocidad de símbolo	facebook	20	GBD
Número de niveles de señal	l	4	
Muestras por UI	METRO	32	
Tasa de error del detector de objetivos	DER0	10-8	
Tensión de salida del transmisor, víctima	AV	0,4	V
Ganancia CC CTLE	gdc	[-12:1:0]	dB
CTLE ganancia CC2	gDC_HP	[-6:1:0]	dB
Polo CTLE HP	fHP_PZ	0,25	GHz
CTLE cero	fZ	fb/2,5	GHz
poste CTLE1	fP1	fb/2,5	GHz
poste CTLE2	fP2	fb	
Cursor principal FFE	c (0)	0,62	
Precursor de FFE	c (-1)	[-0.18:0.02:0]	
Postcursor FFE	c (1)	[-0,38:0,02:0]	
Longitud del DFE	Nótese bien	12	mentos de usuario
Límite de magnitud DFE	bmáx(1)	0,75	
	bmáx(2 Nb)	0,2	
Umbral de paso COM	th	3	dB

Según la configuración adoptada de TX FFE + RX CTLE + RX DFE, la predicción Los parámetros incluyen coeficientes de derivación del FFE, ganancia de CC del CTLE y coeficientes de derivación. del DFE. Los datos TX FFE y RX CTLE se obtienen buscando valores FOM en todo el espacio de diseño, mientras que los coeficientes DFE se determinan de acuerdo con el respuesta de pulso h(t) correspondiente al FOM óptimo. El SI se ve significativamente afectado por poscursores más cercanos al cursor principal, mientras que los poscursores más distantes ejercen un mínimo influencia, por lo que solo se predicen los primeros cuatro poscursores del DFE.

3.3. División de conjuntos de datos

Para este trabajo se recogieron un total de 325 canales. En este conjunto de datos, 280 canales fueron Se utilizó para crear un conjunto de datos de diagrama de ojo, denominado Conjunto de datos A, y se utilizaron 215 canales. para crear un conjunto de datos para los COM y predicar los parámetros de los ecualizadores, referidos como conjunto de datos B. Debido al impacto del FOM, el enfoque COM utiliza diferentes ecualizadores métodos de optimización de parámetros para sistemas stripline y microstrip, lo que resulta en

atasets : : ROM
Conjunto de entrenamiento 184 140
Conjunto de validación 40 40
Conjunto de prueba 56 40 (5 del conjunto de validación) 325
Total

capacidad de ajuste debilitada del modelo ML. En consecuencia, estandarizamos el conjunto de datos B. usando canales strip-line. Entre los 215 canales del conjunto de datos B, 170 canales son stripline conjunto de datos A, mientras que 45 canales son canales de línea de banda recién creados. De acuerdo a ambos conjuntos de datos se dividen en entrenamiento, validación y prueba. del conjunto de pruebas del conjunto de datos B se extrae del conjunto de validación. La configuración dividida

Tabla 3. Configuración de división del conjunto de datos.

Conjuntos de datos	número de canales	
	A: 280	B: 215 (170 de A)
Conjunto de entrenamiento	184	140
Conjunto de validación	40	40
Conjunto de prueba	56	40 (5 del conjunto de validación)
Total	325	

4. Construcción del modelo en cascada y entrenamiento

4.1. DNN

El modelo MP, propuesto por el psicólogo McCulloch y el lógico Pitts, es un modelo matemático que se desarrolló a través de un análisis y síntesis de la metodología estándar adecuada, de neuronas [31]. Este modelo es crucial para la implementación de redes neuronales y constituye la unidad fundamental de una DNN. La estructura de este modelo se representa en Figura 6. La $x = (x_1, x_2, \cdots, x_n)$ representa las n características de entrada de la neurona, y $\omega_i = (\omega_{i1}, \omega_{i2}, \cdots, \omega_{in})$ denota el vector de peso responsable de las conexiones lineales ponderadas. El vector de peso ω se optimiza dentro de las redes neuronales para mejorar la precisión de las predicciones. La salida ponderada lineal Z se puede expresar de la siguiente manera:

$$Z_i = \sum_{j=1}^{n} \omega_{ij} x_j \quad (3)$$

DNN utilizado en este documento, como se muestra en la Figura 7, adopta una estructura estándar con múltiples capas ocultas, junto con una capa de entrada y una capa de salida de regresión lineal [33], y Emplea una arquitectura de red completamente conectada entre sus neuronas.

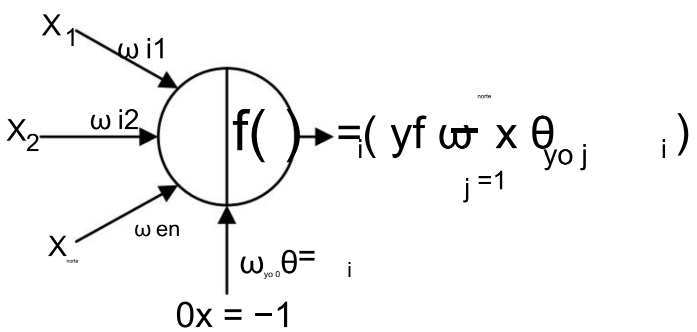


Figura 6. La estructura del modelo de neuronas MP.

La Figura 6 muestra la estructura del modelo de neurona MP. La diferencia entre Z y se utiliza un desplazamiento θ como entrada para la función f(·). La función f(·) se conoce como función de activación y se utiliza para derivar la salida de la neurona. θ puede considerarse como el peso de la entrada x0 con un valor fijo de −1 y también es un objetivo de optimización del sistema neuronal red. En este artículo, la función ReLU se adopta como función de activación porque puede mitigar más eficazmente la aparición de sobreajuste. La fórmula para esto se muestra en Ecuación (4).

$$f_{ReLU}(x) = \text{máx}(0, x) \quad (4)$$

La salida de la función de activación se utiliza normalmente como entrada para capas adyacentes. Las DNN pueden completar tareas como regresión, clasificación y reconocimiento. Los DNN son ampliamente utilizado para tareas de regresión en producción y evaluación del desempeño [4,6–9,32]. El

DNN utilizado en este documento, como se muestra en la Figura 7, adopta una estructura estándar con múltiples capas ocultas, junto con una capa de entrada y una capa de salida de regresión lineal [33], y emplea una arquitectura de red completamente conectada entre sus neuronas.

Figura 6. La estructura del modelo de neuronas MP.

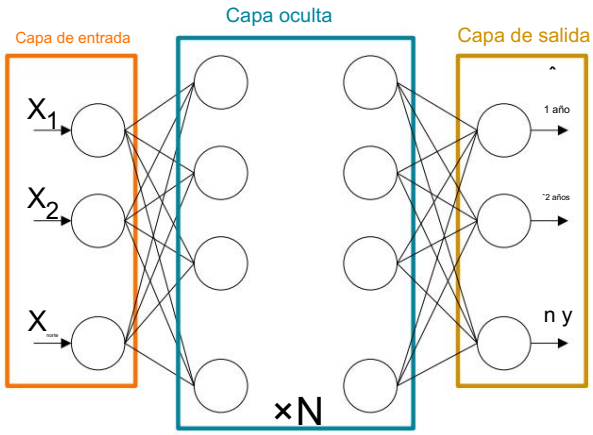


Figura 7. La estructura de la DNN de este estudio.

Para que el modelo converja mejor, el modelo de entrenamiento generalmente requiere una función para evaluar la brecha entre los valores predichos por el modelo y los valores reales. En esto el papel, Smooth L1 [34] se adopta como función de costo:

$$\text{suaveL1} = \begin{cases} 0{,}5\,(yn - \hat{y}n)^2 & \text{si } |yn - \hat{y}n| < 1 \\ |yn - \hat{y}n| - 0{,}5 & \text{en caso contrario} \end{cases}, \quad (5)$$

donde yn y $\hat{y}n$ representan el valor real y el valor predicho, respectivamente.

El desempeño del modelo se evalúa utilizando la función de costos; luego, se emplea un optimizador para propagar hacia atrás y minimizar continuamente la función de costos. Los parámetros y pesos del modelo se actualizan continuamente hasta alcanzar el rendimiento esperado. Después de comparar varios optimizadores diferentes, finalmente se optó por el optimizador Adam, seleccionado para su uso en este artículo [35].

4.2. Transformador para regresión

La precisión de los parámetros del ecualizador y los valores COM predichos directamente por los parámetros de características del canal es relativamente baja. En este trabajo, la respuesta de un solo bit (SBR) antes y después de la ecualización se predice primero en función de los parámetros de características del canal, y luego los parámetros del ecualizador y los valores COM se determinan de acuerdo con la respuesta previa al pulso. La respuesta del pulso debe utilizarse como serie temporal para la predicción.

Tradicionalmente, los modelos RNN o LSTM se utilizan para la predicción de series temporales, pero estos modelos tienen ciertas limitaciones. El modelo Transformer puede mejorar efectivamente el gradiente. Las obras tienen ciertas limitaciones. El modelo Transformer puede mejorar efectivamente el problema de explosión gradual que ocurre en LSTM mediante el uso de mecanismos de atención, mejorando así precisión de predicción [36].

La fórmula de un mecanismo de atención se muestra en la ecuación (6). Se compone principalmente de Q (Consulta), K (Clave) y V (Valor), junto con un módulo softmax. Cuando Q, K, de tres entradas: Q (Query), K (Key) y V (Value), junto con un módulo softmax. Cuando Q y V son idénticos, se denomina mecanismo de autoatención. K y V son idénticos, lo que se denomina mecanismo de autoatención.

$$\text{Atención}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{dk}}\right)V \quad (6)$$

La arquitectura completa de Transformer incluye principalmente una capa de incrustación, una capa de codificación posicional, una capa codificadora, una capa decodificadora y una capa de salida. En esto capa de codificación opcional, una capa codificadora, una capa decodificadora y una capa de salida. En este trabajo, trabajo, dado que nuestra tarea es una tarea de predicción de series de tiempo en lugar de una tarea de PNL, el posicionamiento eliminado la capa de codificación posicional. La esencia de la predicción de series de tiempo puede verse como Se ha eliminado la capa de codificación. La esencia de la predicción de series temporales puede verse como una tarea de regresión, el modelo solo necesita generar un único valor de salida correspondiente a cada punto de tiempo y no necesita un decodificador para generar salidas de secuencia. Usando solo

a cada punto de tiempo y no necesita un decodificador para generar salidas de secuencia. Usando la capa del codificador es suficiente, lo que simplifica la estructura del modelo y reduce la com- solo la capa del codificador carga putacional. Como se muestra en la Figura 8, nuestro diseño final mantiene la capa de incrustación, la capa codificadora y la capa de salida final. Además, para lograr una capa de conversión-codificador más rápida y la capa de salida final. Además, para lograr una conversión más rápida, la capa codificadora y la capa de salida final. Además, para lograr una mayor rapidez

Por ejemplo, se eligió la unidad lineal de error gaussiano (GELU) como función de activación. Por ejemplo, se eligió la Unidad Lineal de Error Gaussiano (GELU) como función de activación. convergencia, se eligió la Unidad Lineal de Error

$$GELU(x) = xP(X \le x) = x\Phi(x) = x \cdot \frac{1}{2}[1 + \operatorname{erf}(x/\sqrt{2})] \quad (7)$$

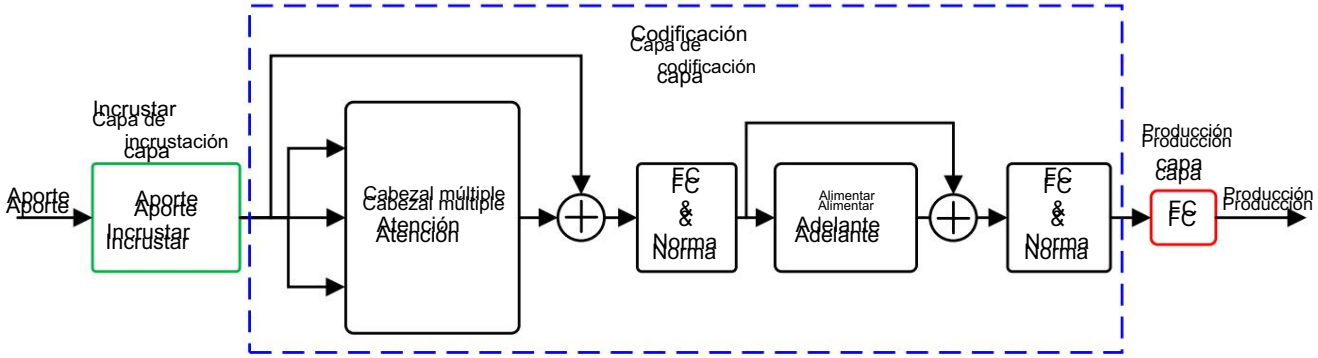


Figura 8. La estructura de nuestro Transformer para tareas de regresión.
Figura 8. La estructura de nuestro Transformer para tareas de regresión.
Figura 8. La estructura de nuestro Transformer para tareas de regresión.

4.3. Modelo en Cascada y Entrenamiento
4.3. Modelo en cascada y entrenamiento

Como se ilustra en la Figura 9, en este caso se utiliza un transformador en cascada y un modelo DNN. Como se ilustra en la Figura 9, en este caso se utiliza un transformador en cascada y un modelo DNN trabajar para predecir con precisión el valor COM y sus correspondientes coeficientes del ecualizador. Un para predecir con precisión el valor COM y sus correspondientes coeficientes del ecualizador. Un trabajo Se han recopilado un total de 325 características de canal como un conjunto de datos y se utiliza un modelo DNN. Se han recopilado un total de 325 características de canal como un conjunto de datos. un modelo DNN es para predecir la característica espectral para cada canal, específicamente el IL y RL en Nyquist usado para predecir la característica espectral para cada canal, específicamente el IL y RL en Nyquist. Estos datos espectrales predichos, junto con las características del canal, son entonces puntos de puntos de frecuencia. Estos datos espectrales predichos, junto con las características del canal, son puntos de frecuencia de Nyquist. introducido en un modelo de transformador para derivar el SBR de preecualización. Utilizar la alimentación preigualada en un modelo de transformador para derivar el SBR de preecualización. Utilizando la precodificación SBR, los parámetros del bajo la configuración COM son ecualización SBR, los parámetros del ecualizador correspondientes bajo la configuración COM ecualizador correspondientes bajo la configuración COM son ización SBR, los parámetros del ecualizador correspondientes calculado por el modelo DNN. Para calcular el SBR posterior a la ecualización, el ecualizador se calcula mediante el modelo DNN. Para calcular el SBR posterior a la ecualización, el modelo DNN calcula el ecualizador. Para calcular el SBR posterior Los parámetros combinados con las características del canal y las características en el dominio de la frecuencia son entonces **los parámetros combinados con las características del canal y las características en el dominio de la frecuencia son entonces** enviado a otro modelo de Transformer. Finalmente, el canal SBR posterior a la ecualización muestreado se envía a otro modelo de transformador para obtener el SBR posterior a la ecualización muestreado se envía a otro modelo de Las características y las características espectrales se usan para calcular el valor COM correspondiente final en las características y características espectrales se utilizan para calcular el valor COM final correspondiente en el modelo DNN, el en el modelo DNN.

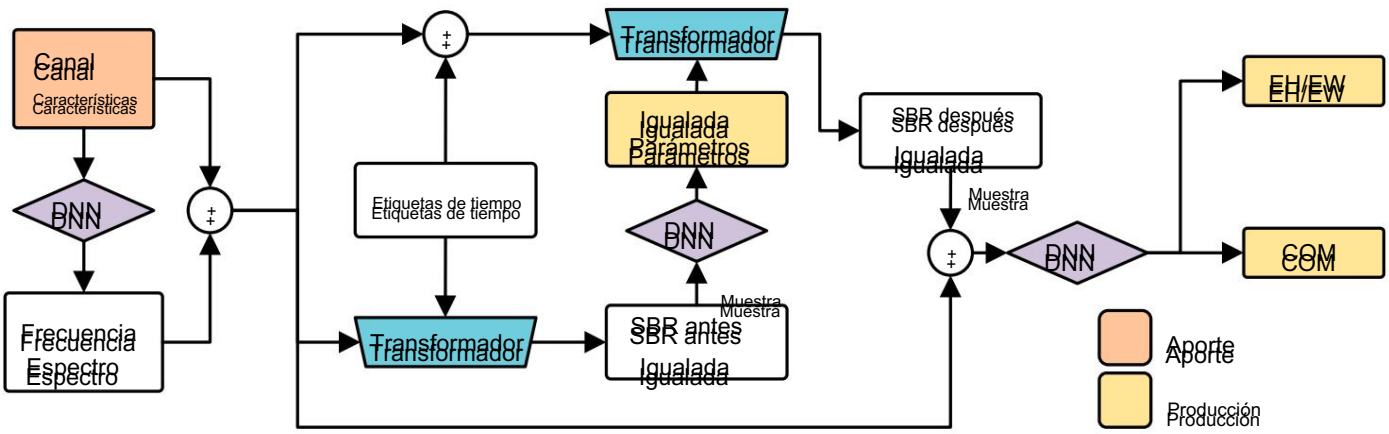


Figura 9. La estructura del modelo en cascada.
Figura 9. La estructura del modelo en cascada.
Figura 9. La estructura del modelo en cascada.

El vector de entrada de características físicas para cada canal se puede definir de la siguiente manera:
El vector de entrada de características físicas para cada canal se puede definir de la siguiente manera:
El vector de entrada de características físicas para cada canal se puede definir de la siguiente manera:

mínimos: (8)

$$X = xij\ i=1:NC, j=1:NP,$$

donde NC representa la cantidad de canales en el conjunto de datos, NP es la cantidad total de características del canal y X es una matriz NC × NP . El elemento xij denota el valor cuantificado de una característica física específica de un canal: por ejemplo, x11 corresponde al valor de la característica de longitud del primer canal.

El DNN utilizado en el modelo en cascada se puede representar mediante la Ecuación (9):

$$\begin{aligned} Y &= FDNN(Z, \theta) \\ &= W(l)\, fReLU(\cdots(W(2)\, fReLU(W(1)X + \theta^{(1)}) + \theta^{(2)})\cdots) + \theta^{(l)} \end{aligned}, \quad (9)$$

donde Z representa la salida ponderada linealmente de la matriz de entrada X y las ponderaciones del modelo W, como se describe en la Ecuación (3), y θ representa el desplazamiento. El índice l denota la capa del modelo DNN, que incluye la capa de entrada, las capas ocultas y la capa de salida. Nuestra red emplea un total de cuatro modelos DNN diferentes con diferentes hiperparámetros: Y_f^{mi} para predecir características espectrales, Y_C^{mj} para predecir parámetros del ecualizador, Y para predecir el diagrama de ojo e Y para predecir COM.

La matriz aumentada Xf se obtiene concatenando horizontalmente la matriz de entrada X del primer modelo DNN con su salida Y_f :

$$\begin{aligned} Xf &= [\, XY_f \,] \\ &= \begin{matrix} x11 & x12 & \cdots & x1NP & y1I & y1R & \cdots & x2NP & y2I \\ x21 & x22 & & & & y2R & & & \\ \vdots & \vdots & \ddots & \vdots & & \vdots & & \vdots & \\ xNC1 & xNC2 & \cdots & xNCNP & yNC\,I & & yNCR & & \end{matrix} \end{aligned}, \quad (10)$$

donde el vector columna $yI = (y1I, y2I, \cdots, yNc\,I)^t$ representa el canal predicho IL y $yR = (y1R, y2R, \cdots, yNcR)^t$ representa el canal previsto RL.

Para mejorar la precisión del valor COM previsto, la respuesta del pulso se extrae del SBR previo y posterior a la ecualización en la herramienta COM. El punto de tiempo correspondiente al pico se toma como cursor principal, y las amplitudes de cuatro precursores y ocho postcursores determinados por el intervalo de muestreo COM se seleccionan como el SBR requerido. Luego, el pulso de cada canal se aplana y se insertan las etiquetas de tiempo correspondientes, formando un vector de características secuencial. La ecuación (11) representa la secuencia de valores de amplitud. Después de este procesamiento, se puede emplear el modelo Transformer para predecir con precisión el SBR.

$$\begin{aligned} SBR1(t) &= FTrans[x11(t), x12(t), \cdots, x1NP(t), y1I(t), y1R(t)] \\ SBR1 &= [SBR1(t - t0), \cdots, SBR1(t), \cdots, SBR1(t + t1)] \end{aligned}, \quad (11)$$

donde FTrans denota el modelo de transformador que se utiliza, SBR1(t) representa el SBR correspondiente al vector de características de entrada del canal en un punto de tiempo específico, y SBR1 representa la secuencia SBR del canal. t0 y t1 son los límites izquierdo y derecho del rango de tiempo, respectivamente.

Al introducir el SBR en el modelo DNN posterior con distintos parámetros, la predicción de los parámetros del ecualizador se puede lograr mediante la Ecuación (12):

$$Y_C^{mj} = FDNN(SBR1, SBR2, \cdots, SBRNC) \quad (12)$$

Al integrar los parámetros del ecualizador con la entrada que presenta características espectrales, Y_f^{mi}, se construye para la predicción de la post-ecualización.

SBR. Esta matriz, a través de la Ecuación (13), luego se procesa en una secuencia dependiente del tiempo siguiendo la metodología presentada en la Ecuación (11).

$$\begin{aligned} SBReq\_\_\_\_(t) &= FTrans[xNC1(t), \cdots, xNCNP(t), \\ &\qquad yNCI(t), yNCR(t), Y^{mi}\_\_\_\_(t)] \\ SBReq\_\_\_\_ &= [SBReq\_\_\_\_(t-t0), \cdots, \\ &\qquad SBReq\_\_\_\_(t), \cdots (t+tf)] \\ Y^{mi}\_\_\_\_ Grifos(t) &= [y(t), y(t), y(t)]^{DCganar}_{FFE\text{-}NC\ CTLE\text{-}NC\ DFE\text{-}NC} \end{aligned} \tag{13}$$

Finalmente, el SBR posterior a la ecualización obtenido se concatena con X f para formar el resultado final. características de entrada para predecir el valor COM.

Antes de entrenar formalmente el modelo en cascada, estandarizamos los datos de entrada y f req empleó estrategia de preentrenamiento. Inicialmente, los modelos DNN F DNN(X) y F una DNN(X) diseñados para predecir características espectrales y parámetros del ecualizador, respectivamente, se entrenaron de forma independiente y se guardaron sus parámetros correspondientes. Estos modelos DNN previamente entrenados se integraron con el modelo Transformer necesario para construir el modelo en cascada completo para predecir los parámetros del ecualizador y COM, como se muestra en la Figura 9. Para entrenar el modelo en cascada, se usó Smooth L1 como función de costo. y se aplicó el algoritmo de Adam para actualizar los parámetros del modelo. Con la ayuda del kit de herramientas de Optuna, la estructuración del modelo y la optimización de hiperparámetros (HPO) se realizaron mediante optimización bayesiana utilizando el GP [37,38]. La función de adquisición elegida fue la función de Mejora Esperada (EI), con un conjunto inicial de 10 puntos de observación, lo que permite una HPO efectiva dentro de un alcance computacionalmente factible.

En comparación con los métodos de búsqueda de cuadrícula global, búsqueda de cuadrícula aleatoria y búsqueda de reducción a la mitad, la optimización bayesiana es más eficiente para HPO y requiere menos tiempo de optimización. Además , en este artículo se emplea el método de validación cruzada K-Fold para mejorar la confiabilidad de la evaluación del modelo.

Después de completar el entrenamiento del modelo, sus capacidades predictivas se pueden evaluar con varias métricas diferentes. Una de estas métricas, que se muestra en la ecuación (14), es el RMSE.
La operación de cuadratura en RMSE aumenta su sensibilidad a los errores, lo que la hace particularmente sensible a los valores atípicos. Esta mayor sensibilidad también puede hacer que responda demasiado a los valores atípicos, pero la operación de raíz cuadrada no puede reflejar intuitivamente la magnitud real del error.

$$RMSE = \sqrt{\frac{1}{n}\sum_{yo=1}^{n}(yn - \hat{y}n)^2} \tag{14}$$

MAE se calcula según el valor promedio de las diferencias absolutas entre los valores previstos y reales. En comparación con RMSE, MAE es menos sensible a los valores atípicos y proporciona un reflejo más intuitivo de la magnitud real de los errores. Sin embargo, esto puede llevar a una subestimación del impacto de los errores.

$$MAE = \frac{1}{n}\sum_{yo=1}^{n}|yn - \hat{y}n| \tag{15}$$

MAPE es el valor promedio de las diferencias porcentuales absolutas entre los valores previstos y reales. En comparación con RMSE y MAE, MAPE es más fácil de entender y permite comparar resultados de predicción en diferentes escalas. Es más adecuado para situaciones con cambios significativos en los valores reales. Sin embargo, cuando los valores reales son cercanos a cero, el error puede volverse grande, por lo que MAPE no es adecuado para conjuntos de datos que contienen valores cero o cercanos a cero.

$$MAPA = \frac{1}{n}\sum_{yo=1}^{n}\frac{yn - \hat{y}n}{\sum yn} \tag{dieciséis}$$

valores cero o cercanos a cero.

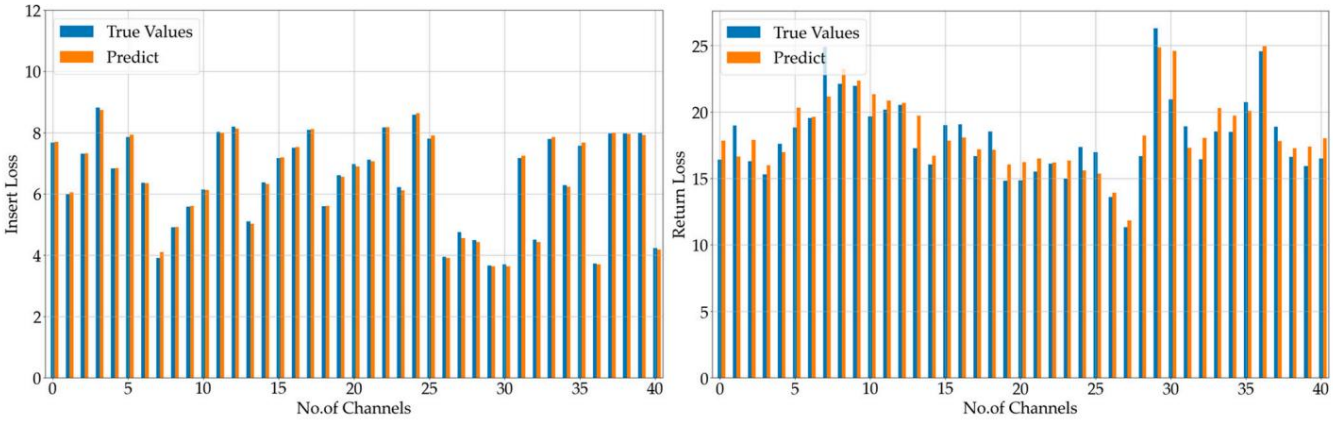
$$MAPA = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (16)$$

5. Resultados numéricos

En este trabajo, el modelo de red neuronal se ha construido utilizando una biblioteca Python 5. Resultados numéricos
Basado en Pytorch 2.0.1. El ordenador utilizado para el cálculo del modelo incluía una GeForce. En este trabajo, el
RTX 4070 y un Intel Core i5-12600 KF 3,7 GHz. Basado en el modelo de red neuronal se ha construido utilizando una biblioteca Python.
Pytorch 2.0.1. La computadora utilizada para el cálculo del modelo incluía una GeForce
RTX 4070 y un Intel Core i5-12600 KF 3,7 GHz.

5.1. IL y RL

Como se muestra en la Figura 10, IL y RL son pérdidas en la frecuencia de Nyquist y el modelo 5.1. IL y RL
Las predicciones de IL son más precisas que sus predicciones de RL. Esto se debe a que IL es primordial.
Como se muestra en la Figura 10, IL y RL son pérdidas en la frecuencia de Nyquist, y el modelo
Muy influenciados por la pérdida dieléctrica y las características de entrada, incluidas las predicciones de IL relacionadas
con la pérdida dieléctrica, son más precisas que sus predicciones de RL. Esto se debe a que la IL es principalmente
parámetros. Por el contrario, RL se ve afectado por factores adicionales como la impedancia, mal influenciada por la
pérdida dieléctrica y las características de entrada, incluidos los parámetros relacionados con la pérdida dieléctrica y las
formas complejas de las trazas, lo que complica su predicción. Por tanto, la precisión de los eters. Por el contrario, RL se
ve afectado por factores adicionales como desajustes de impedancia y
La predicción IL es mejor que la predicción RL. El error relativo de la predicción de RL son las formas de trazas complejas,
lo que complica su predicción. Por lo tanto, la precisión de la predicción de IL
17,4%, pero esto es suficiente para respaldar la predicción precisa de los diagramas de ojo y COM es mejor que la
predicción de RL. El error relativo de la predicción de RL es del 17,4%, pero esto es
valores.
suficiente para respaldar la predicción precisa de diagramas de ojo y valores COM.



(a) (b)

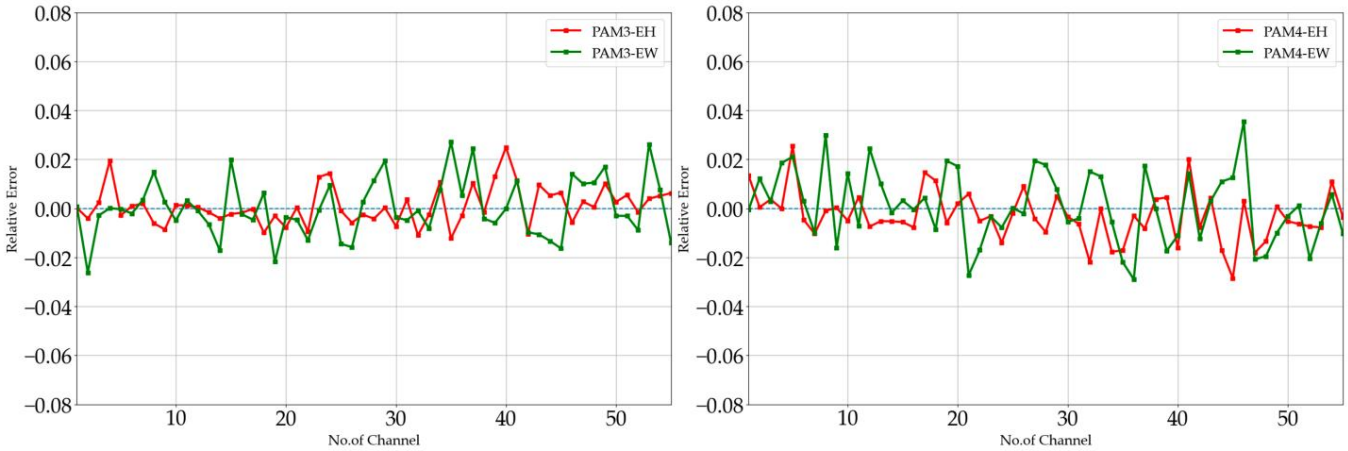
Figura 10. Comparación de los valores previstos y reales para (a) IL y (b) RL.
Figura 10. Comparación de los valores previstos y reales para (a) IL y (b) RL.

5.2. EH/EW de los diagramas del ojo

5.2. EH/EW de los diagramas del ojo

La Figura 11 muestra los errores relativos en la altura y el ancho del ojo para 40 canales de prueba con
La Figura 11 muestra los errores relativos en la altura y el ancho del ojo para 40 canales de prueba con los
los métodos de modulación PAM3 y PAM4. Los canales N° 1 a N° 32 son canales stripline, mientras que los canales N°
métodos de modulación PAM3 y PAM4 los canales No. 1 al No. 32 son canales stripline
33 a N° 56 son canales microstrip. Canales No. 1 al No. 10
neis, mientras que los canales No. 33 al No. 56 son canales microstrip. Los canales No. 1 a No. 10 exhiben trayectorias
exhiben rutas de seguimiento más complejas, lo que resulta en mayores errores en la altura y el ancho de los ojos. Por
de seguimiento más complejas, lo que resulta en mayores errores en la altura y el ancho de los ojos. En
el contrario, los canales No. 11 al No. 23, que tienen un enrutamiento más simple con líneas más rectas o menos
En contraste, los canales No. 11 a No. 23, que tienen un enrutamiento más simple con trayectorias más rectas o menos
curvadas, muestran una mejor precisión de ajuste. La mayoría de los canales del conjunto de datos utilizan FR-4 como
Los caminos de curvatura muestran una mejor precisión de ajuste. La mayoría de los canales en el conjunto de datos
capa dieléctrica, mientras que los canales No. 24 a No. 32 emplean materiales alternativos, lo que resulta en errores
utilizan FR-4 como capa dieléctrica, mientras que los canales No. 24 a No. 32 emplean materiales alternativos, lo que resulta
relativamente mayores. En comparación con los canales stripline, los canales No. 33 al No. 56
en errores relativamente mayores. En comparación con los canales stripline, los canales No. 33 a No. 56, que utilizan
líneas microstrip, exhiben errores relativos significativamente mayores,
que utilizan líneas microstrip, muestran errores relativos significativamente mayores, lo que demuestra que
que las líneas microstrip tienen capacidades antiinterferentes más débiles que las líneas strip. Tabla 4 RMSE, MAE,
las líneas microstrip tienen capacidades antiinterferentes más débiles que las líneas strip. La tabla 4 muestra la
MAPE y MRE para la predicción del diagrama de ojo con PAM3 y PAM4
muestra RMSE, MAE, MAPE y MRE para la predicción del diagrama de ojo con los métodos de modulación PAM3. Es
evidente que los resultados de la predicción exhiben una buena convergencia.
y métodos de modulación PAM4. Es evidente que los resultados de la predicción exhiben buena precisión y precisión.
Debido a que la unidad EH se elige como mV, las métricas RMSE y MAE aumentan
vergencia y precisión. Debido a que la unidad EH se elige como mV, las métricas RMSE y MAE en un orden de magnitud
en comparación con EW. Los resultados de esta tabla, combinados con los
aumento en un orden de magnitud en comparación con EW. Los resultados de esta tabla,
combinados en la Figura 11, demuestran que los modelos PAM3 y PAM4 son efectivos y exhiben
alta precisión.

Ojo	Modulación	RMSE	MAE	MAPA	ERM (%)
EH (mV)	PAM3	4,8 × 10-1	3,3 × 10−1	5,7 × 10−3	2.5
	PAM4	4,0 × 10-1	2,9 × 10−1	8,1 × 10−3	2.8
EW (interfaz de usuario)	PAM3	4,4 × 10−3	3,4 × 10−3	9,3 × 10−3	2,7
	PAM4	4,4 × 10−3	3,6 × 10−3	1,2 × 10−2	3,5



(a) PAM3 (b) PAM4

Figura 11. Errores relativos de los diagramas de ojo para los canales del conjunto de prueba en los dos formatos de modulación.

Figura 11. Errores relativos de los diagramas de ojo para los canales del conjunto de prueba en los dos formatos de modulación.

5.3. COM

Tabla 4. Predicción de diagramas de ojo con el modelo DNN.

Ojo	Modulación	RMSE	MAE	MAPA	ERM (%)
EH (mV)	PAM3	4,8 × 10-1	3,3 × 10−1	5,7 × 10−3	2.5
	PAM4	4,0 × 10-1	2,9 × 10−1	8,1 × 10−3	2.8
EW (interfaz de usuario)	PAM3	4,4 × 10−3	3,4 × 10−3	9,3 × 10−3	2.7
	PAM4	4,4 × 10−3	3,6 × 10−3	1,2 × 10−2	3.5

Debido a que muchos valores de SBR previos y posteriores al cursor son cercanos o iguales a cero, Modulation Eye MAPE y MRE no son adecuados para evaluar la precisión de las predicciones de SBR, y solo RMSE y MAE se utilizan para evaluar la precisión de la predicción de SBR. Para el SBR 7,3 × 10−3 antes de la ecualización de los 40 canales de prueba, el RMSE es 7,3 × 10−4 y el MAE 8,1 × 10−3 Para el SBR después de la ecualización, el RMSE es 1,1 × 10−3 y el MAE es 1,6 × 10−4. A continuación, utilizamos un seleccionado aleatoriamente (ID: 3) del conjunto de prueba para ilustrar el rendimiento de la predicción SBR 4,4 × 10−3 3,6 × 10−3 PAM4 . Como se muestra en la Figura 12, el modelo Transformer puede predecir con alta precisión tanto la SBR previa como la posterior a la ecualización.

5.3. COM

Debido a que muchos valores SBR previos y posteriores al cursor son cercanos o iguales a cero, MAPE y MRE no son adecuados para evaluar la precisión de las predicciones SBR, y solo RMSE y MAE se utilizan para evaluar la precisión de la predicción del SBR. Para el SBR antes ecualización de los 40 canales de prueba, el RMSE es 7,3 × 10−4 y el MAE es 6,3 × 10−4 . Para el SBR después de la ecualización, el RMSE es 1,1 × 10−3 y el MAE es 1,6 × 10−4 . A continuación, nosotros Utilice un canal seleccionado aleatoriamente (ID: 3) del conjunto de prueba para ilustrar la predicción de SBR. actuación. Como se muestra en la Figura 12, tanto la preecualización como la postecualización El modelo Transformer puede predecir el SBR con alta precisión.

Para el formato de señal PAM4, la Figura 13 proporciona una demostración concisa. Seleccionamos una configuración de ecualizador que incluye un TX FFE de 3 derivaciones, un RX CTLE con doble ganancia de CC, y un RX DFE de 12 grifos (en lo sucesivo denominado configuración de ecualizador estándar). Como Como se muestra en la Figura 13, los canales No. 7 y No. 30 tienen grandes errores relativos debido a su formas complejas. Algunos canales con RL alto, como los canales No. 7–No. 9, también exhiben error relativo significativo. El error relativo de los valores COM previstos para cada prueba. El canal está dentro del 2%, lo que demuestra la capacidad del modelo DNN-Transformer para lograr la precisión de predicción requerida.

Aquí, las influencias de diferentes combinaciones de ecualizadores en la precisión de la predicción. de los valores COM se analizan en profundidad. Como se muestra en la Tabla 5, configuramos cinco diferentes combinaciones de ecualizador. Los resultados indican que nuestro modelo en cascada propuesto logra el rendimiento de predicción esperado en varias combinaciones de ecualizador. Sin embargo lo és Es evidente que una reducción en la variedad de ecualizadores debilita la generalización del modelo. capacidad y disminuye su precisión de predicción. Cuando solo estaba habilitado el DFE, el RMSE aumentado a 1,6 × 10−1, y el MRE alcanzó el 17,5%. Atribuimos esto a la optimización. método de este ecualizador dentro del marco COM y el hecho de que el modelo en cascada También tiene en cuenta las predicciones tanto para el CTLE como para el FFE. Por lo tanto, ajustamos la estructura de la red eliminando ecualizadores no utilizados en el modelo en cascada para diferentes

combinaciones de ecualizador. Esta modificación ha supuesto una mejora en el modelo.
precisión de predicción para los valores COM, con el MRE disminuyendo del 17,5% al 11,5%.

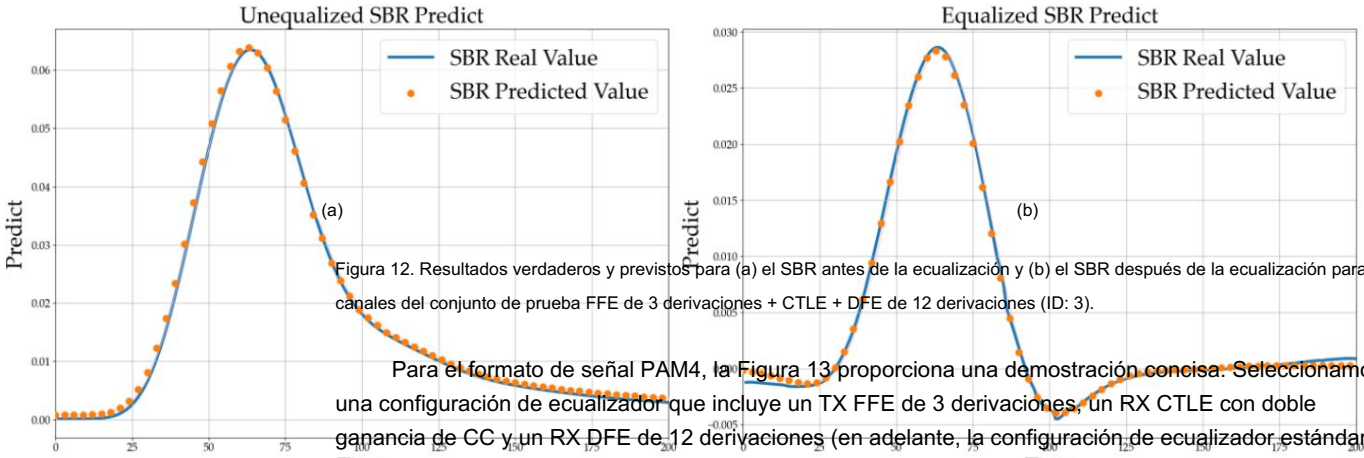


Figura 12. Resultados verdaderos y previstos para (a) el SBR antes de la ecualización y (b) el SBR después de la ecualización para canales del conjunto de prueba FFE de 3 derivaciones + CTLE + DFE de 12 derivaciones (ID: 3).

Para el formato de señal PAM4, la Figura 13 proporciona una demostración concisa. Seleccionamos una configuración de ecualizador que incluye un TX FFE de 3 derivaciones, un RX CTLE con doble ganancia de CC y un RX DFE de 12 derivaciones (en adelante, la configuración de ecualizador estándar). Como se muestra en la Figura 13, los canales No. 7 y No. 30 tienen grandes errores relativos debido a sus formas complejas. Algunos canales con RL alto, como los canales No. 7–No. 9, también exhiben un error relativo significativo. El error relativo de los valores COM pronosticados para cada canal de prueba está dentro del 2%, lo que demuestra la capacidad del modelo DNN-Transformer para lograr la precisión de predicción requerida.

(b)

Resultados verdaderos y previstos para (a) el SBR antes de la ecualización y (b) el SBR después de la ecualización para canales del equipo de prueba FFE de 3 derivaciones + CTLE + DFE de 12 derivaciones (ID: 3).

Para el formato de señal PAM4, la Figura 13 proporciona una demostración concisa. Seleccionamos una configuración de ecualizador que incluye un TX FFE de 3 derivaciones, un RX CTLE con ganancia de CC dual y un RX DFE de 12 derivaciones (en lo sucesivo denominado configuración de ecualizador estándar). Como se muestra en la Figura 13, los canales No. 7 y No. 30 tienen grandes errores relativos debido a sus formas complejas. Algunos canales con RL alto, como los canales No. 7–No. 9, también exhiben un error relativo significativo. El error relativo de los valores COM pronosticados para cada canal de prueba está dentro del 2%, lo que demuestra la capacidad del modelo DNN-Transformer para lograr la precisión de predicción requerida.

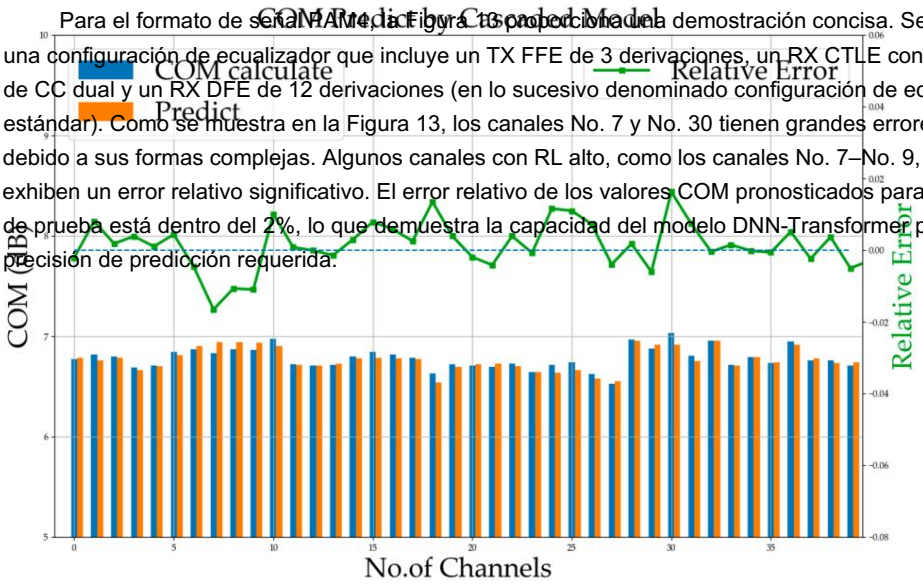


Figura 13. Error relativo y valores COM para la codificación PAM4 (ID: 1–40) en la configuración del ecualizador estándar.

Figura 13. Error relativo y valores COM para la codificación PAM4 (ID: 1–40) en el ecualizador estándar configuración.

Aquí, analizamos en profundidad las influencias de diferentes combinaciones de ecualizadores en la precisión de predicción de los valores COM pronosticados. Como se muestra en la Tabla 5, configuramos cinco combinaciones de ecualizador diferentes. Los resultados indican que nuestro modelo en cascada propuesto logra el rendimiento de predicción esperado en varias combinaciones de ecualizadores. Sin embargo, es evidente que una reducción en la variedad de ecualizadores debilita la capacidad del modelo.

La variación de los valores COM se analiza en profundidad. Como se muestra en la Tabla 5, configuramos cinco combinaciones de ecualizador. Los resultados indican que nuestro modelo en cascada propuesto logra el rendimiento de predicción esperado en varias combinaciones de ecualizadores. Sin embargo, es evidente que una reducción en la variedad de ecualizadores debilita las

Tabla 5. Valores de error COM para diferentes combinaciones de ecualizadores.

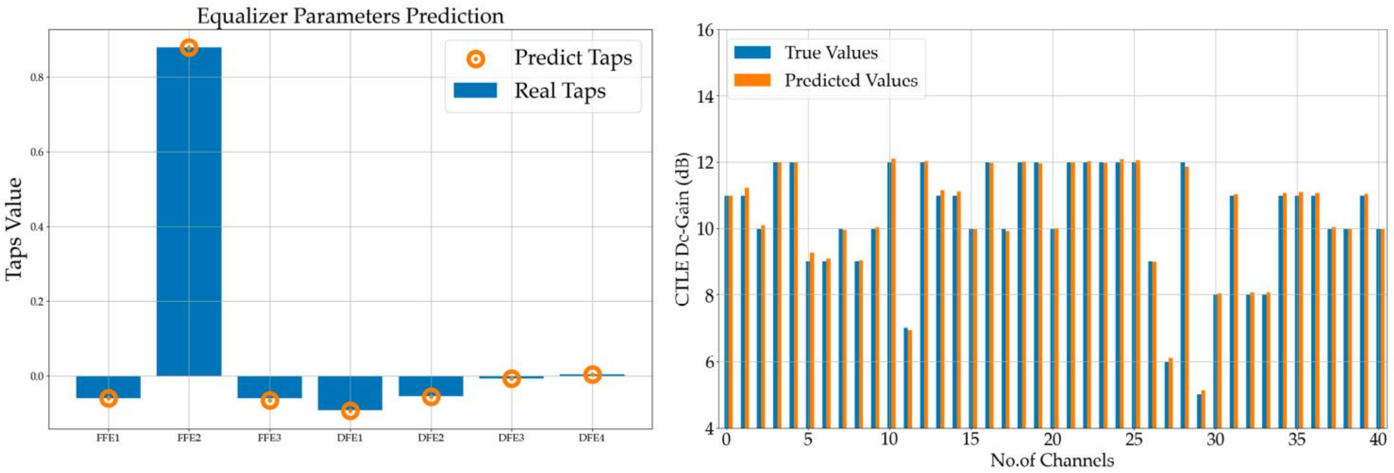
COM	RMSE	MAE	MAPE	MRE
FFE de 3 grifos + CTLE + DFE de 12 grifos	5,0 × 10−2	3,7 × 10−2	5,6 × 10−3	1,6%
FFE de 3 derivaciones + CTLE + DFE de 8 derivaciones	4,2 × 10−2	3,2 × 10−2	4,1 × 10−3	2,0%
FFE de 3 derivaciones + CTLE + DFE de 4 derivaciones	9,6 × 10−2	7,1 × 10−2	1,4 × 10−2	1,7%
Configuración CTLE + DFE de 12 derivaciones	1,6 × 10-1	1,1 × 10-1	3,0 × 10-2	5,7%
DFE de 12 grifos				17,5%

Aquí se analizan en profundidad las influencias de diferentes combinaciones de ecualizadores en la precisión de predicción de los valores COM. Como se muestra en la Tabla 5, configuramos cinco combinaciones de ecualizador diferentes. Los resultados indican que nuestro modelo en cascada propuesto logra el rendimiento de predicción esperado en varias combinaciones de ecualizadores. Sin embargo, es evidente que una reducción en la variedad de ecualizadores debilita la capacidad del modelo.

3 tomas FFE + CTLE + 12 tomas DFE	4,5 × 10−2	3,5 × 10−2	5,1 × 10−3	1,6%
3 tomas FFE + CTLE + 8 tomas DFE	5,0 × 10−2	3,7 × 10−2	5,6 × 10−3	2,0%
3 tomas FFE + CTLE + 4 tomas DFE	4,2 × 10−2	3,2 × 10−2	4,1 × 10−3	1,7%
CTLE + DFE de 12 derivaciones	9,6 × 10−2	7,1 × 10−2	1,4 × 10−2	5,7%
DFE de 12 derivaciones	1,6 × 10−1	1,1 × 10−1	3,0 × 10−2	17,5%

A continuación, se utiliza un canal de prueba (ID: 3) para ilustrar la capacidad de predicción del modelo. con los primeros cuatro grifos del DFE y tres grifos del FFE. Como se muestra en la Figura 14a, el modelo en cascada realiza predicciones muy precisas. Además, los resultados de la predicción para la ganancia de CC principal del CTLE para los 40 canales de prueba, como se presenta en la Figura 14b, también indican un alto nivel de precisión.

A continuación, se utiliza un canal de prueba (ID: 3) para ilustrar la capacidad de predicción del modelo. con los primeros cuatro grifos del DFE y tres grifos del FFE. Como se muestra en la Figura 14a, el predicción realiza predicciones muy precisas. Además, el modelo en cascada de resultados de predicción para la ganancia de CC principal del CTLE para los 40 canales de prueba, como se presenta en la Figura 14b, también indican un alto nivel de precisión.



(a) Grifos FFE y DFE

(b) Ganancia CC de CTLE

Figura 14. Comparación de los resultados de los parámetros del ecualizador reales y previstos.

Figura 14. Comparación de los resultados de los parámetros del ecualizador reales y previstos.

Se investigaron las capacidades predictivas de diferentes combinaciones de modelos en cascada. Se cerrado de la siguiente manera. La configuración del ecualizador se estableció como configuración estándar y los parámetros de HPO se configuraron de manera uniforme, con el rango de búsqueda de tasa de aprendizaje de 0,0001 a 0,01, realizando 20 pruebas para su optimización. Dadas las características del SBR, probamos el rendimiento predictivo de modelos independientes, incluido un DNN, y probamos el Red LSTM y un transformador. El DNN se combina con dos redes de predicción de secuencias y un LSTM (DNN-LSTM) y un transformador, se seleccionaron como un esquema en cascada. Se seleccionó el El modelo Transformer en la Figura 9 y conforme a la red LSTM típica. Reemplaza el DNN, LSTM y Transformer independientes sin módulos de procesamiento de series de tiempo, por lo que estos modelos solo pueden lograr predicciones de valores COM y diagramas de ojo. El mejor parámetros estructurales de estos modelos, como el número de capas ocultas y el tamaño de la capa oculta del LSTM, y el número de cabezales, la dimensión del vector y el tamaño de la capa oculta de dimensión de red del mecanismo de atención del Transformador, se obtuvieron todos a través Optimización bayesiana. Como se presenta en la Tabla 6, los modelos en cascada demuestran una calidad superior. capacidades predictivas en comparación con los modelos independientes. Entre los modelos en cascada, aquellos que incorporan un transformador muestran el mejor rendimiento predictivo, lo que indica que El modelo en cascada DNN-Transformer es el modelo de predicción COM óptimo.

Tabla 6. Valores de error previstos de diferentes modelos.

Modelo	RMSE	MAE	MAPA	ERM (%)
DNN	6,9 × 10−2	4,3 × 10−2	7,8 × 10−3	6.5
LSTM	5,3 × 10−2	4,4 × 10-2	6,4 × 10−3	2.0
DNN-LSTM	5,0 × 10−2	4,1 × 10-2	6,0 × 10−3	1.8
Transformador	4,7 × 10−2	3,7 × 10−2	5,5 × 10−3	1.7
DNN–Transformador	4,6 × 10−2	3,5 × 10−2	5,1 × 10−3	1.6

Como se muestra en la Tabla 7, hemos comparado el modelo DNN-Transformer con los modelos presentado en la sección Introducción en términos de funcionalidad. El LSTM tradicional [20]

La estructura no pudo lograr la predicción bajo variables de parámetros físicos. Tradicional
Los modelos de aprendizaje automático, como los modelos SVM [13], RFR [12] y DNN [9] , tienen limitaciones.
capacidades de procesamiento de secuencias y, por lo tanto, no pueden predecir formas de onda transitorias. El
El modelo GNN-RNN [21] simplifica los componentes y circuitos interconectados, es decir
que algunos de los parámetros reales de los enlaces se pierdan y el rendimiento disminuya. El
Los modelos anteriores no pudieron lograr la optimización del enlace y RFR [12] solo pudo completar FD.
predicción. Por el contrario, el DNN-Transformer propuesto es más práctico. Este modelo puede
logra efectivamente estas tareas de comportamiento anteriores y se desempeña mejor que otros modelos
en RMSE y MAE para la predicción del valor COM. Debido a la creciente complejidad de este
modelo, su tiempo de inferencia aumenta, pero aún tiene una ventaja de tiempo significativa en comparación
a la simulación EM tradicional cuyo tiempo de solución única es de varias horas.

Tabla 7. Comparativa funcional de diferentes modelos para enlaces de alta velocidad.

Modelo	Físico Parámetros Variable	Transitorio Forma de onda Predicción	FD Predicción	Enlace Mejoramiento	GPU/CPU Tiempo (ms)	RMSE	MAE
SVM [13]	Sí	No	No	No	<5 *	3,8 × 10−1	3,1 × 10−1
RFR [12]	Sí	No	Sí	No	<5 *	4,8 × 10-1	3,8 × 10−1
DNN [9]	Sí	No	No	No	<5	6,9 × 10−2	4,3 × 10−2
LTM [20]	No	Sí	No	No	<5	5,3 × 10−2	4,4 × 10−2
GNN-RNN [21]	Sí	Sí	No	No	567.1		\
DNN–Transformador	Sí	Sí	Sí	Sí	792,8 *	\ 4,6 × 10-2	3,5 × 10−2

* Tiempo de inferencia de GPU.

6. Conclusiones

En este artículo, se propone una red neuronal en cascada DNN-Transformer para
Análisis del SI en enlaces serie de alta velocidad. Durante el proceso de creación del conjunto de datos, hicimos
referencia a los estándares USB4 Gen4 y 50GBASE-KR para diseño de PCB y electromagnética.
simulación, y utilizó los parámetros de diseño físico de cada canal como entradas para el
modelo. Este modelo DNN-Transformer se utiliza en este artículo para extraer las características de
parámetros de diseño físico de canales y predecir con éxito los datos del diagrama de ojo
y valores COM de enlaces de prueba. Además, este modelo de aprendizaje profundo puede lograr con éxito
predecir el SBR antes y después de la ecualización, y los parámetros principales del ecualizador para diferentes
Las combinaciones también se predicen con precisión. Para el entrenamiento del modelo, empleamos un
Método de optimización bayesiano basado en el GP para HPO. Finalmente, este artículo compara
DNN–Transformer con varios otros modelos, como los modelos DNN, LSTM, Transformer y DNN-LSTM. Los
resultados muestran que nuestro modelo en cascada DNN-Transformer
logra la predicción del rendimiento y la optimización de la arquitectura de ecualización para alta velocidad
enlaces seriales y el MRE en sus resultados de predicción COM para el conjunto de prueba, con un ecualizador
configuración que comprende un TX FFE de 3 tomas, un RX CTLE con doble ganancia de CC y un 12 tomas
RX DFE, es del 1,6%.

Contribuciones de los autores: Conceptualización, LW, JZ e YZ (Yinhang Zhang); metodología, LW,
JZ y HJ; software, LW; validación, LW, JZ e YZ (Yinhang Zhang); análisis formal, LW,
HJ y YZ (Yinhang Zhang); investigación, LW; recursos, LW, YZ (Yinhang Zhang) y
YZ (Yongzheng Zhan); curación de datos, LW e YZ (Yinhang Zhang); escritura—borrador original
preparación, LW; redacción: revisión y edición, XY, YZ (Yinhang Zhang) e YZ (Yongzheng
Zhan); visualización, LW; supervisión, XY, YZ (Yinhang Zhang) e YZ (Yongzheng Zhan);
administración de proyectos, LW; adquisición de financiación, LW, YZ (Yinhang Zhang) y XY Todos los autores
He leído y aceptado la versión publicada del manuscrito.

Financiamiento: Este trabajo fue apoyado por la Fundación Nacional de Ciencias Naturales de China (Subvención No.
61861019, 62161012), el Departamento de Educación Provincial de Hunan (Subvención No. 22B0525, 21A0335),

la Fundación de Ciencias Postdoctorales de China (Subvención No. 2024M751268) y la Fundación de Investigación de Postgrado Programa de la Universidad de Jishou (Subvención No. Jdy23035).

Declaración de disponibilidad de datos: Los datos utilizados para respaldar los hallazgos del estudio están disponibles en el artículo.

Conflictos de intereses: el autor Yongzheng Zhan era empleado de la empresa Shandong Yunhai Guochuang Cloud Computing Equipment Industry Innovation Company Limited. El restante Los autores declaran que la investigación se realizó sin ningún propósito comercial o financiero. relaciones que podrían interpretarse como un potencial conflicto de intereses.

Abreviaturas

En este manuscrito se utilizan las siguientes abreviaturas:

DNN	Red neuronal profunda;
LSTM	Red neuronal de memoria a corto plazo;
SBR	Respuesta de un solo bit;
SI	Integridad de la señal
COM	margen operativo del canal;
eCOM	mejoró el margen operativo del canal;
GP	proceso gaussiano;
HPO	Optimización de hiperparámetros;
EM	electromagnético;
EMFS	solucionador de campos electromagnéticos;
IBIS-AMI	Interfaz de modelo algorítmico de especificación de información del búfer de entrada/salida;
EH	altura de los ojos;
EW	ancho de los ojos;
rl	pérdida de retorno;
IL	pérdida de inserción;
DT	dominio del tiempo;
FD	dominio de la frecuencia;
ml	aprendizaje automático;
SVM	Máquina de vectores de soporte;
LS-SVM	Máquina de vectores de soporte de mínimos cuadrados;
FFN	Red neuronal de avance;
RFR	Regresión de bosque aleatorio;
RNN	Red Neuronal Recurrente;
FFE	ecualizador anticipado;
DFE	ecualizador de retroalimentación de decisiones;
CTLE	ecualizador lineal de tiempo continuo;
ILD	desviación de la pérdida de inserción;
ICR	relación entre pérdida de inserción y diafonía;
ISI	interferencia entre símbolos;
FOM	figura de mérito.

Referencias

1. Salón, SH; Heck, HL Integridad de señal avanzada para diseños digitales de alta velocidad; John Wiley & Sons: Hoboken, Nueva Jersey, EE. UU., 2009; págs. 201–206.
2. Yan, J.; Zargaran-Yazd, A. Modelado IBIS-AMI de interfaces de memoria de alta velocidad. En las actas del IEEE 24th Electrical de 2015 Performance of Electronic Packaging and Systems (EPEPS), San José, CA, EE. UU., 25 a 28 de octubre de 2015; págs. 73–76.
3. Comité de Estándares LAN/MAN de la IEEE Computer Society. Estándar IEEE para Ethernet. 2012. Disponible en línea: https://ieeexplore.ieee.org/document/6419735 (consultado el 26 de junio de 2024).
4. Wang, Y.; Hu, QS Optimización de enlaces serie de alta velocidad basada en COM mediante aprendizaje automático. IEICE Trans. Electrón. 2022, 105, 684–691. [Referencia cruzada]
5. Gore, B.; Richard, M. Un ejercicio de aplicación del margen operativo del canal (COM) para el diseño de canales 10GBASE-KR. En procedimientos del Simposio Internacional IEEE sobre Compatibilidad Electromagnética (EMC) de 2014, Raleigh, Carolina del Norte, EE. UU., 4 a 8 de agosto de 2014; págs. 653–658.

6. Ambasana, N.; Gope, B.; Mutnury, B.; Anand, G. Aplicación de redes neuronales artificiales para la predicción de la altura y el ancho de los ojos a partir de parámetros S. En actas de la 23ª Conferencia del IEEE sobre rendimiento eléctrico de sistemas y embalajes electrónicos, Portland, Oregón, EE. UU., 26 a 29 de octubre de 2014; págs. 99-102.

7. Ambasana, N.; Anand, G.; Mutnury, B.; Gope, D. Predicción de la altura/ancho de los ojos a partir de parámetros S utilizando modelos basados en el aprendizaje. Traducción IEEE. Componente. Paquete. Fabricante. Tecnología. 2016, 6, 873–885. [Referencia cruzada]
8. Ambasana, N.; Anand, G.; Gope, D.; Mutnury, B. Método de identificación de frecuencia y parámetro S para análisis ocular basado en ANN Predicción de altura/ancho. Traducción IEEE. Componente. Paquete. Fabricante. Tecnología. 2017, 7, 698–709. [Referencia cruzada]
9. Lu, T.; Sol, J.; Wu, K.; Yang, Z. Modelado de canales de alta velocidad con métodos de aprendizaje automático para el análisis de la integridad de la señal. Traducción IEEE. Electromagn. Compat. 2018, 60, 1957–1964. [Referencia cruzada]
10. Lho, D.; Parque, J.; Parque, H.; Kang, H.; Parque, S.; Kim, J. Método de estimación del ancho y la altura de los ojos basado en una red neuronal artificial (ANN) para USB 3.0. En actas de la 27.ª Conferencia del IEEE de 2018 sobre rendimiento eléctrico de sistemas y envases electrónicos (EPEPS), San José, CA, EE. UU., 14 a 17 de octubre de 2018; págs. 209-211.

11. Sánchez-Masís, A.; Rimolo-Donadio, R.; Roy, K.; Sulaimán, M.; Schuster, C. Modelos FNN para la regresión de parámetros S en interconexiones multicapa con diferentes longitudes eléctricas. En Actas de la Conferencia Latinoamericana de Microondas (LAMC) IEEE MTT-S 2023, San José, CA, EE. UU., 6 al 8 de diciembre de 2023; págs. 82–85.

12. Li, X.; Hu, Q. Modelado de canales basado en aprendizaje automático para enlaces serie de alta velocidad. En actas de la sexta conferencia internacional sobre informática y comunicaciones (ICCC) del IEEE de 2020, Chengdu, China, 11 a 14 de diciembre de 2020; págs. 1511-1515.
13. Trinchero, R.; Canavero, FG Modelado de la altura del diagrama de ojo en enlaces de alta velocidad mediante máquina de vectores de soporte. En actas del 22.º taller del IEEE de 2018 sobre integridad de la señal y la energía (SPI), Brest, Francia, 22 a 25 de mayo de 2018; págs. 1–4.
14. Cao, Y.; Zhang, QJ Un nuevo enfoque de entrenamiento para el modelado robusto de redes neuronales recurrentes de circuitos no lineales. Traducción IEEE . Microondas. Teoría Tecnológica. 2009, 57, 1539–1553. [Referencia cruzada]
15. Mutnury, B.; Swaminathan, M.; Libous, JP Macromodelado de controladores de E/S digitales no lineales. Traducción IEEE. Adv. Paquete. 2006, 29, 102–113. [Referencia cruzada]
16. Cao, Y.; Erdin, I.; Zhang, QJ Modelado de comportamiento transitorio de controladores de E/S no lineales que combinan redes neuronales y circuitos equivalentes. Microondas IEEE. Alambre. Componente. Letón. 2010, 20, 645–647. [Referencia cruzada]
17. Yu, H.; Michalka, T.; Larbi, M.; Swaminathan, M. Modelado del comportamiento de controladores de E/S sintonizables con énfasis previo que incluye Ruido de la fuente de alimentación. Traducción IEEE. Integral de muy gran escala. (VLSI) Sistema. 2020, 28, 233–242. [Referencia cruzada]
18. Yu, H.; Chalamalasetty, H.; Swaminathan, M. Modelado de osciladores controlados por voltaje, incluido el comportamiento de E/S mediante Redes neuronales aumentadas. Acceso IEEE 2019, 7, 38973–38982. [Referencia cruzada]
19. Faraji, A.; Noohi, M.; Sadrossadat, SA; Mirvakili, A.; Na, WC; Feng, F. Red neuronal recurrente profunda normalizada por lotes para macromodelado de circuitos no lineales de alta velocidad. Traducción IEEE. Microondas. Teoría Tecnológica. 2022, 70, 4857–4868. [Referencia cruzada]
20. Moradi, M.; Sadrossadat, A.; Derhami, V. Redes neuronales de memoria a corto plazo para modelar electrónica no lineal Componentes. Traducción IEEE. Componente. Paquete. Fabricante. Tecnología. 2021, 1, 840–847. [Referencia cruzada]
21. Li, Z.; Li, CX; Wu, ZM; Zhu, Y.; Mao, JF Modelado sustituto de enlaces de alta velocidad basados en GNN y RNN para la integridad de la señal Aplicaciones. Traducción IEEE. Microondas. Teoría Tecnológica. 2023, 71, 3784–7796. [Referencia cruzada]
22. Lho, D.; Parque, H.; Parque, S.; Kim, S.; Kang, H.; Sim, B.; Kim, S.; Parque, J.; Cho, K.; Canción, J.; et al. Modelos de redes neuronales profundas basadas en características de canal para una estimación precisa del diagrama de ojo en un intercalador de silicio con memoria de alto ancho de banda (HBM). Traducción IEEE. Electromagn. Compat. 2021, 64, 196–208. [Referencia cruzada]
23. Li, GS; Mao, CS; Zhao, modelo de regresión semisupervisado de WS para la estimación del diagrama de ojo del intercalador de silicio con memoria de alto ancho de banda (HBM). En las actas del Simposio de la Sociedad Internacional de Electromagnetismo Computacional Aplicado de 2023, Hangzhou, China, del 15 al 18 de agosto de 2023; págs. 1–3.

24. Goay, CH; Ahmad, NS; Goh, P. Redes convolucionales temporales para simulación transitoria de canales de alta velocidad. Alex. Ing. J. 2023, 74, 643–663. [Referencia cruzada]
25. Li, BW; Jiao, B.; Chou, CH; Mayder, R.; Franzon, P. Modelo de aprendizaje profundo en cascada de autoevolución para receptores de alta velocidad Adaptación. Traducción IEEE. Componente. Paquete. Fabricante. Tecnología. 2020, 10, 1043–1053. [Referencia cruzada]
26. Li, BW; Jiao, B.; Chou, CH; Mayder, R.; Franzon, P. Adaptación de CTLE mediante aprendizaje profundo en enlaces SerDes de alta velocidad. En Actas de la 70.ª Conferencia de Tecnología y Componentes Electrónicos (ECTC) de IEEE, Orlando, FL, EE. UU., 3 a 30 de junio de 2020; págs. 952–955.

27. Zhang, HH; Xue, ZS; Liu, XY; Li, P.; Jiang, LJ; Shi, GM Optimización del canal de alta velocidad para la integridad de la señal con Deep Algoritmo genético. Traducción IEEE. Electromagn. Compat. 2022, 64, 1270–1274. [Referencia cruzada]
28. Shan, G.; Li, G.; Wang, Y.; Xing, C.; Zheng, Y.; Yang, Y. Aplicación y perspectiva de métodos de inteligencia artificial en la predicción de la integridad de la señal y optimización de microsistemas. Micromáquinas 2023, 14, 344. [CrossRef]
29. Mellitz, R.; Corrió, A.; Li, diputado; Ragavassamy, V. Margen operativo del canal (COM): evolución de las especificaciones del canal para 25 Gbps y más. En Actas de DesignCon 2013, Santa Clara, CA, EE. UU., 28 a 31 de enero de 2013.
30. Pu, B.; Él, J.; Armon, A.; Guo, Y.; Liu, Y.; Cai, Q. Metodología de diseño de integridad de señal para paquetes en ópticas co-empaquetadas basada en la figura de mérito como margen operativo del canal. En actas del Simposio internacional conjunto EMC/SI/PI y EMC Europe de IEEE de 2021, Raleigh, Carolina del Norte, EE. UU., 26 de julio a 13 de agosto de 2021; págs. 492–497.
31. McCulloch, WS; Pitts, W. Un cálculo lógico de las ideas inmanentes a la actividad nerviosa. Toro. Matemáticas. Biofísica. 1943, 5, 115-133. [Referencia cruzada]

32. Bono, FM; Radicioni, L.; Cinquemani, S. Un enfoque novedoso para el control de calidad de líneas de producción automatizadas que funcionan bajo Condiciones altamente inconsistentes. Ing. Aplica. Artif. Intel. 2023, 122, 106149. [Referencia cruzada]
33. Acero, RGD; Torrie, JH Principios y procedimientos de estadística; McGraw Hill: Nueva York, Nueva Jersey, Estados Unidos, 1960.
34. Girshick, R. Rápido R-CNN. En Actas de la Conferencia Internacional IEEE sobre Visión por Computadora (ICCV), Santiago, Chile, 7–13 diciembre de 2015; págs. 1440-1448.
35. Kingma, DP; Ba, J. Adam: un método de optimización estocástica. arXiv 2014, arXiv:1412.6980.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gómez, AN; Káiser, L.; Polosukhin, I. La atención es todo lo que necesita. En Actas de los avances en sistemas de procesamiento de información neuronal, Long Beach, CA, EE. UU., 4 a 9 de diciembre de 2017; págs. 6000–6010.
37. Akiba, T.; Sanó, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: un marco de optimización de hiperparámetros de próxima generación. En actas de la 25.ª Conferencia internacional ACM SIGKDD sobre descubrimiento de conocimientos y minería de datos, Anchorage, Alaska, EE. UU., 4 a 8 de agosto de 2019.
38. Frazier, PI Un tutorial sobre optimización bayesiana. arXiv 2018, arXiv:1807.02811.

electrónica

MDPI

Artículo

Mejora de la percepción de los vehículos autónomos en condiciones climáticas adversas:

Un modelo de objetivos múltiples para la clasificación integrada del clima y la detección de objetos

Nasser Aloufi *, Abdulaziz Alnori y Abdullah Basuhail

Departamento de Ciencias de la Computación, Facultad de Computación y Tecnología de la Información, Universidad Rey Abdulaziz, Jeddah 21589, Arabia Saudita; asalnori@kau.edu.sa (AA); abasuhail@kau.edu.sa (AB)
* Correspondencia: nhassanaloufi@stu.kau.edu.sa

Resumen: La detección sólida de objetos y la clasificación climática son esenciales para el funcionamiento seguro de vehículos autónomos (AV) en condiciones climáticas adversas. Si bien las investigaciones existentes a menudo tratan estas tareas por separado, este artículo propone un novedoso modelo de objetivos múltiples que trata la clasificación del clima y la detección de objetos como un problema único utilizando únicamente el sistema de detección de cámara AV. Nuestro modelo ofrece una mayor eficiencia y posibles ganancias de rendimiento al integrar la evaluación de la calidad de la imagen, la Red Generativa Adversarial de Superresolución (SRGAN) y una versión modificada de You Only Look Once (YOLO), versión 5. Además, al aprovechar la desafiante Detección en condiciones climáticas adversas Nature (DAWN), que incluye cuatro tipos de condiciones climáticas severas, incluido el clima arenoso que a menudo se pasa por alto, hemos aplicado varias técnicas de aumento, lo que resultó en una expansión significativa del conjunto de datos de 1027 imágenes a 2046 imágenes. Además, optimizamos la arquitectura YOLO para una detección sólida de seis clases de objetos (automóvil, ciclista, peatón, motocicleta, autobús, camión) en escenarios climáticos adversos. Experimentos integrales demuestran la efectividad de nuestro enfoque, logrando una precisión promedio promedio (mAP) del 74,6 %, lo que subraya el potencial de este modelo de objetivos múltiples para mejorar significativamente las capacidades de percepción de las cámaras de los vehículos autónomos en entornos desafiantes.

Palabras clave: vehículos autónomos; red neuronal convencional; detección de objetos; aprendizaje profundo; sensores de cámara; clima adverso; clasificación del clima

Cita: Aloufi, N.; Alnori, A.; Basuhail, A. Mejora de la percepción de vehículos autónomos en condiciones climáticas adversas: un modelo de objetivos múltiples para la clasificación integrada del clima y la detección de objetos. Electrónica 2024, 13, 3063. https://doi.org/10.3390/electrónica13153063

Editores académicos: Shiping Wen y Jianying Xiao

Recibido: 14 de mayo de 2024
Revisado: 22 de julio de 2024
Aceptado: 29 de julio de 2024
Publicado: 2 de agosto de 2024

1. Introducción

El rápido avance de la tecnología de vehículos autónomos (AV) ha captado la atención de investigadores, ingenieros, formuladores de políticas y el público. Un elemento central del desarrollo AV son los sensores que permiten la percepción y la toma de decisiones dentro de entornos de conducción dinámicos. Entre ellos, los sensores de las cámaras desempeñan un papel vital como fuente principal de percepción visual en los sistemas AV. Las cámaras capturan imágenes de alta resolución en tiempo real de los alrededores del vehículo, proporcionando datos visuales cruciales para la detección y clasificación precisa de diversos objetos. Al aprovechar algoritmos avanzados de detección de objetos, las cámaras contribuyen a diversas funcionalidades AV, como el mantenimiento de carril y la planificación de rutas, al monitorear continuamente las marcas de carril y los cambios en el trazado de la carretera. Esto permite que el vehículo mantenga su posición dentro de los carriles y tome decisiones informadas sobre la trayectoria y las maniobras, mejorando así la seguridad vial general y el flujo de tráfico. Además, los sensores de las cámaras contribuyen a la planificación de rutas identificando obstáculos, señales de tráfico y otras entidades, lo que permite que el vehículo adapte su trayectoria en consecuencia y navegue por escenarios de tráfico complejos.

La estimación de la profundidad es otra capacidad clave de los sensores de las cámaras, que permite a los vehículos autónomos percibir con precisión las distancias de los objetos circundantes. A través de técnicas avanzadas de procesamiento de imágenes , las cámaras pueden proporcionar percepción de profundidad, mejorando la conciencia espacial del vehículo y las capacidades para evitar obstáculos. La segmentación de imágenes es una tarea adicional realizada

Capacidad de concienciación y evitación de obstáculos. La segmentación de imágenes es una tarea adicional que realizan los sensores de las cámaras, en la que la escena visual se segmenta en regiones semánticamente significativas . Esta segmentación permite al AV distinguir entre varios elementos dentro de su campo de visión, facilitando una sólida detección y clasificación de objetos, esencial para una navegación segura. Además, los sensores de las cámaras desempeñan un papel crucial en la fusión de la percepción, integrando datos de múltiples cámaras ubicadas alrededor del vehículo para construir una visión completa. Esta fusión mejora la comprensión del vehículo de su entorno, permitiéndole tomar decisiones informadas en tiempo real. Además de externos, los sensores de las cámaras también contribuyen al seguimiento del habitáculo y proporcionan datos esenciales del comportamiento y el estado de los pasajeros. La Figura 1 enumera las funciones principales de las cámaras y sus sistemas en AV.

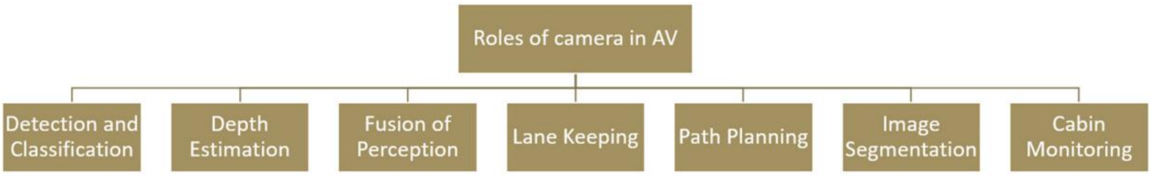


Figura 1. Funciones principales de las cámaras y sus sistemas en los AV.

Además de sus funciones principales, las cámaras ofrecen una solución rentable y ligera en comparación con tecnologías de sensores alternativas como LiDAR y radar. Esta asequibilidad facilita la adopción y el despliegue generalizados de la tecnología AV, lo que allana el camino para un futuro en el que los vehículos autónomos sean omnipresentes en nuestras carreteras.

Dicho esto sobre las funciones de las cámaras, alcanzar el nivel más alto de AV objetos, la "automatización total" [1], requiere que los vehículos detecten todos los en todas las condiciones y escenarios con nada menos que un aspecto humano. comportamiento. Alcanzar este nivel sigue siendo un gran desafío durante condiciones climáticas adversas, que requiere imágenes confiables del medio ambiente. Condiciones climáticas como lluvia, nieve, niebla y tormentas de arena claras plantea desafíos importantes para los sensores de las cámaras, y los desafíos clave incluyen:

- Visibilidad reducida: las condiciones climáticas adversas a menudo conducen a una visibilidad reducida, la lluvia, la nieve y la niebla pueden oscurecer el campo de visión, lo que dificulta que las cámaras distingan objetos y obstáculos con precisión.
- Gotas de agua y acumulación de nieve: la lluvia y la nieve pueden provocar gotas de agua o acumulación de nieve en las lentes de las cámaras, lo que provoca distorsión, borrosidad u oclusión de imágenes capturadas. Esta acumulación puede degradar la calidad de la imagen y dificultar la capacidades de detección y reconocimiento de objetos.
- Niebla y neblina: las condiciones de niebla crean una atmósfera nebulosa que reduce el contraste y compromete la confiabilidad de las imágenes de la cámara, dificultando la detección y localización de objetos. La presencia de la niebla y la neblina dificultan que las cámaras distingan los objetos del fondo, comprometiendo la confiabilidad de los sistemas de percepción AV.
- Deslumbramiento y reflejos: el deslumbramiento de las superficies de la carretera mojadas o las superficies reflectantes puede causar reflejos en las imágenes de la cámara, lo que da como resultado imágenes sobreexpuestas o descoloridas. El deslumbramiento y los reflejos pueden oscurecer información visual importante, lo que dificulta la tarea de AV para navegar con seguridad en condiciones climáticas adversas.
- Partículas de arena y polvo: las tormentas de arena y las condiciones de polvo pueden provocar la acumulación de partículas de arena y polvo en las lentes de las cámaras obstruyen el campo de visión y se degradan. Esta acumulación de partículas puede comprometer el rendimiento de los sensores de las cámaras, lo que afecta la confiabilidad de los sistemas de percepción AV.
- Condiciones de iluminación dinámicas: las condiciones climáticas adversas pueden provocar cambios rápidos en las condiciones de iluminación, incluidas variaciones en el brillo, el contraste y la temperatura del color. Los sistemas de cámaras deben adaptarse a estas condiciones de iluminación dinámica para mantener imágenes precisas y garantizar una detección y reconocimiento fiables de objetos.

Los sistemas de cámaras deben adaptarse a estas condiciones de iluminación dinámica para mantener una percepción precisa del entorno y garantizar una detección y reconocimiento fiables de objetos.

- Calibración del sensor: las condiciones climáticas adversas pueden requerir ajustes en la cámara. • parámetros de calibración para compensar cambios en iluminación, visibilidad y sensor por parámetros de calibración para compensar cambios en iluminación, visibilidad y sensor rendimiento. Garantizar una calibración precisa del sensor es esencial para mantener el confiable. Garantizar una calibración precisa del sensor es esencial para mantener el Fiabilidad y eficacia de los sensores de las cámaras en condiciones climáticas adversas. idad y eficacia de los sensores de las cámaras en condiciones climáticas adversas.
- Fiabilidad y robustez: las condiciones meteorológicas adversas suponen fiabilidad y robustez Fiabilidad y robustez las condiciones meteorológicas adversas suponen fiabilidad y robustez desafíos para los sensores de las cámaras, que requieren que continúen funcionando de manera efectiva desafíos para los sensores de las cámaras, que requieren que continúen funcionando de manera efectiva en duras condiciones ambientales. Garantizar la durabilidad y resistencia de las levas en condiciones ambientales adversas. Garantizar la durabilidad y resistencia de la cámara

Los sensores de la cámara para el reconocimiento seguro y confiable de los vehículos autónomos en condiciones climáticas adversas. Los sensores de la cámara para el reconocimiento seguro y confiable de los vehículos autónomos en condiciones climáticas adversas.

La Figura 2 muestra algunos desafíos de detección de objetos en condiciones climáticas adversas. Cada tipo de La Figura 2 muestra algunos desafíos de detección de objetos en condiciones climáticas adversas. Cada tipo de El clima tiene sus propios obstáculos. Por ejemplo, en caso de fuertes nevadas y tormentas de arena, la carretera El clima tiene sus propios obstáculos. Por ejemplo, en caso de fuertes nevadas y tormentas de arena, la carretera la cámara puede quedar cubierta por la nieve y la arena. Durante la lluvia, las gotas de agua en la cámara La cámara puede quedar cubierta por la nieve y la arena. Durante la lluvia, las gotas de agua en la cámara Las lentes provocan distorsión y borrosidad en las imágenes capturadas. Todos los desafíos mencionados Las lentes provocan distorsión y borrosidad en las imágenes capturadas. Todos los desafíos mencionados anteriormente dificultan que la cámara perciba objetos con precisión. Los desafíos dificultan que la cámara perciba objetos con precisión.

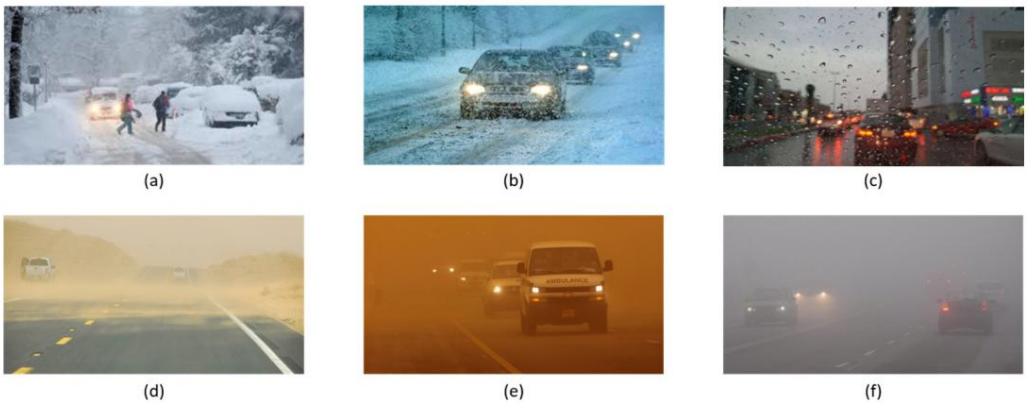


Figura 2. Algunos desafíos climáticos adversos. (a,b) Las fuertes nevadas pueden oscurecer los carriles. (c) La lluvia conduce Figura 2. Algunos desafíos climáticos adversos. (a,b) Las fuertes nevadas pueden oscurecer los carriles. (c) La lluvia provoca distorsión y borrosidad en las imágenes capturadas. (d,e) Una tormenta de arena puede oscurecer los límites de la distorsión y borrosidad en las imágenes capturadas. (d,e) Una tormenta de arena puede oscurecer los límites de la carretera y cambiar la iluminación de la escena. (f) La niebla dificulta que las cámaras distingan objetos de carretera y cambiar la iluminación de la escena. (f) La niebla dificulta que las cámaras distingan objetos de sus antecedentes. sus antecedentes.

Al explorar la detección de objetos utilizando redes neuronales convolucionales (CNN), encontramos Al explorar la detección de objetos utilizando redes neuronales convolucionales (CNN), encontramos dos enfoques principales: el de una etapa y el de dos etapas. El enfoque de dos etapas presenta dos enfoques principales: el de una etapa y el de dos etapas. El enfoque de dos etapas presenta inicialmente la introducción del modelo CNN basado en regiones (R-CNN) en 2014, implica una inicialmente la introducción del modelo CNN basado en regiones (R-CNN) en 2014, implica una etapa de propuesta de región para identificar regiones que contienen objetos, seguida de la extracción de características etapa de propuesta de región para identificar regiones que contienen objetos, seguida de la extracción de características y clasificación de objetos [2]. Sin embargo, este método fue lento debido al procesamiento de cada región propuesta y clasificación de objetos [2]. Sin embargo, este método fue lento debido al procesamiento de cada región propuesta por separado. Fast R-CNN, presentado el año siguiente, mejoró la por separado. Fast R-CNN, introducido el año siguiente, mejoró la velocidad al pasar la velocidad al pasar la imagen completa a través de CNN, generando un mapa de características para la imagen completa a través de CNN, generando un mapa de características para la detección de objetos [3]. Luego se introdujo un R-CNN [4] más rápido para mejorar aún más el rendimiento. detección [3]. Luego se introdujo un R-CNN [4] más rápido para mejorar aún más el rendimiento. A En 2017 se produjo un avance significativo con la introducción de Mask R-CNN [5]. Mascarilla En 2017 se produjo un avance significativo con la introducción de Mask R-CNN [5]. R-CNN adopta Feature Pyramid Network (FPN) como su columna vertebral [6] e introduce una Mask R-CNN adopta Feature Pyramid Network (FPN) como su columna vertebral [6] y una fase innovadora del proceso de detección al generar una máscara de segmentación para cada objeto. introduce una nueva fase en el proceso de detección generando una máscara de segmentación. Por otro lado, el cada objeto. enfoque de una etapa, demostrado por primera vez por Redmon et al. [7] con el modelo YOLO, encapsula todo el proceso de detección en un solo paso Por otro lado, el enfoque de una etapa, demostrado por primera vez por Redmon et al. [7] con el modelo YOLO, que encapsula todo el proceso de detección en un solo paso llamado Darknet-53 para la arquitectura. Una nueva versión de las iteraciones posteriores como YOLOv3 [8] introdujeron un nuevo back- YOLO llamada YOLOv4 entonces a través de la CNN. YOLOv2 y las iteraciones posteriores como YOLOv3 [8] introdujeron una propuesta destinada a mejorar tanto la precisión como la velocidad y lograron una mejora notable. una nueva columna vertebral llamada Darknet-53 para la arquitectura. Una nueva versión de YOLO llamada Luego, YOLOv4 propuso mejorar tanto la precisión como la velocidad y logró un YOLO notable. Otro modelo predecesores. YOLOv5, YOLOv7 [9] y YOLOv8 fueron iteraciones posteriores de llamado Detector multibox de disparo único (SSD), propuesto en [10] mejoras respecto a sus predecesores. YOLOv5, YOLOv7 [9] y YOLOv8 se iteralizaron posteriormente y logró resultados competitivos en el conjunto de datos VOC2007, con mejoras en el mapa. ciones de YOLO. Otro modelo llamado Detector multibox de disparo único (SSD), propuesto en En este artículo, nuestro objetivo es proponer una solución para AV basada en sensores de cámara que [10], ha logrado resultados competitivos en el conjunto de datos VOC2007, con mejoras que no solo pueden detectar objetos, sino también clasificar el clima según la condición de la escena. el mapa.

El alcance de nuestro artículo se muestra en la Figura 3.

En este artículo, nuestro objetivo es proponer una solución para AV basada en sensores de cámara que no solo puedan detectar objetos, sino también clasificar el clima según la condición de la escena. El alcance de nuestro artículo se muestra en la Figura 3.

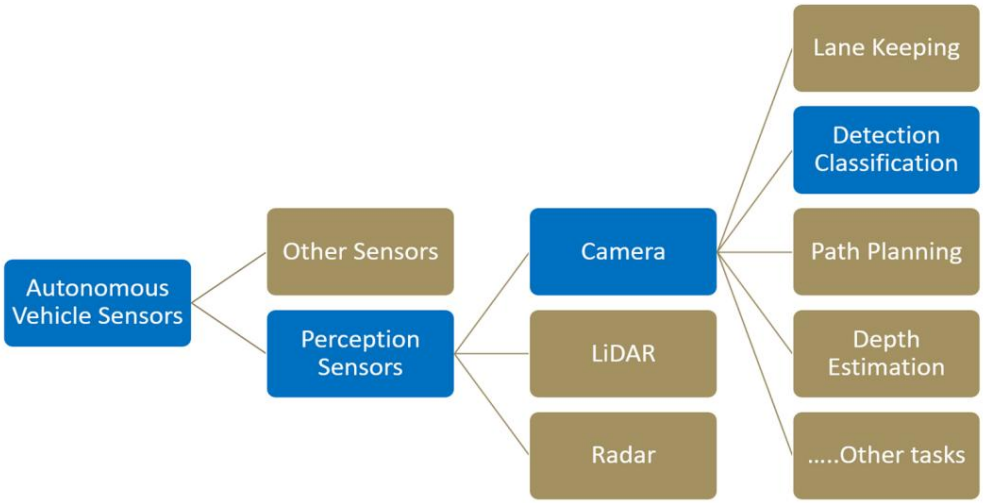


Figura 3. El alcance de este documento está resaltado en azul.

Figura 3. El alcance de este documento está resaltado en azul.

Las principales aportaciones de nuestro trabajo son las siguientes:

Las principales aportaciones de nuestro trabajo son las siguientes:

- Proponemos un modelo multiobjetivos para clasificar el clima y detectar objetos. Como demostraremos en la Sección 2, y hasta donde sabemos, los documentos AV existentes

• Proponemos un modelo multiobjetivos para clasificar el clima y detectar objetos. Como demostraremos en la Sección 2, y hasta donde sabemos, los documentos AV existentes

Lo demostraremos en la Sección 2 y, hasta donde sabemos, los artículos AV existentes tratan la clasificación del clima y la detección de objetos como problemas separados. Nuestra propuesta

Trate la clasificación del clima y la detección de objetos como problemas separados. Nuestro modelo propuesto

trata la clasificación meteorológica de la carretera y la detección de objetos en la carretera como un problema unificado.

El modelo trata la clasificación meteorológica de la carretera y la detección de objetos en la carretera como un problema unificado

para los sistemas de detección de cámaras.

para sistemas de detección de cámaras.

- Hemos ampliado el conjunto de datos de Detección en condiciones meteorológicas adversas en la naturaleza (DAWN) agregando imágenes aumentadas que cubren los cuatro tipos de condiciones climáticas (arenosas, agregando imágenes aumentadas que cubren los cuatro tipos de condiciones climáticas (arenosa, lluviosa, con niebla y nevada). El tamaño total del conjunto de datos casi se ha duplicado, pasando de lluvioso, con niebla y con nieve). El tamaño total del conjunto de datos casi se ha duplicado, aumentando con respecto a su tamaño original.

• Hemos ampliado el conjunto de datos de Detección en condiciones climáticas adversas en la naturaleza (DAWN) agregando imágenes aumentadas que cubren los cuatro tipos de condiciones climáticas (arenosas, agregando imágenes aumentadas que cubren los cuatro tipos de condiciones climáticas (arenosa, lluviosa, con niebla y nevada). El tamaño total del conjunto de datos casi se ha duplicado, pasando de lluvioso, con niebla y con nieve). El tamaño total del conjunto de datos casi se ha duplicado, aumentando con respecto a su tamaño original.

su tamaño original.

- Además de proporcionar un modelo único para clasificar el clima y detectar objetos, abordamos una brecha crítica en la investigación de vehículos autónomos al considerar el clima arenoso.

• Además de proporcionar un modelo único para clasificar el clima y detectar objetos, abordamos una brecha crítica en la investigación de vehículos autónomos al considerar el clima arenoso.

Abordamos una brecha crítica en la investigación de vehículos autónomos considerando las condiciones climáticas arenosas, que los estudios existentes han pasado por alto en gran medida.

condiciones que han sido ignoradas en gran medida por los estudios existentes. • La

arquitectura base del modelo de detección de objetos Solo miras una vez (YOLO)

• La arquitectura base del modelo de detección de objetos You Only Look Once (YOLO) versión 5 ha sido adaptada y modificada para adaptarse a nuestro dominio. Como resultado, tenemos la versión 5 que ha sido adaptada y modificada para adaptarse a nuestro dominio. Como resultado, hemos aumentado con éxito la precisión promedio promedio (mAP) al 74,6%, lo cual es un resultado prometedor.

Aumentó con éxito la precisión promedio media (mAP) al 74,6%, que es un resultado en comparación con otros artículos que utilizaron el mismo conjunto de datos.

resultado prometedor en comparación con otros artículos que utilizaron el mismo conjunto de datos.

2. Trabajo relacionado

2. Trabajo relacionado

Detectar objetos en condiciones climáticas desafiantes presenta dificultades porque el

La detección de objetos en condiciones climáticas desafiantes presenta dificultades porque la calidad de las

imágenes se degrada y las características visuales se ven comprometidas debido a fenómenos climáticos como lluvia, niebla, nieve y tormentas de arena. Estas condiciones afectan el rendimiento de la detección.

La lluvia, la niebla, la nieve y las tormentas de arena reducen la visibilidad de los objetos y complican la diferenciación de los objetos de los elementos circundantes. Se han publicado varios artículos con el objetivo de proponer soluciones adecuadas.

En [11], los autores estudiaron la clasificación del clima adverso junto con el nivel de luz en el

En [11], los autores abordaron el problema de la clasificación del clima adverso junto con el nivel de luz en el

Entorno [11], los autores abordaron el problema de la percepción en condiciones climáticas adversas y poca luz.

Entorno. La degradación de la precisión es una preocupación importante. Los autores introdujeron

Para abordar esta cuestión de la percepción en condiciones climáticas adversas y con poca luz, su propio conjunto de datos. El conjunto de datos fue diseñado para cubrir tres tipos de clima (niebla, lluvia y nieve) y tres niveles de iluminación (brillante, moderada y baja), junto con tres tipos de calles relacionados con el tipo de clima, nivel de iluminación y tipo de calle. Los autores utilizaron ResNet18 como

condiciones, donde la degradación de la precisión es una preocupación importante, los autores introdujeron su propio conjunto de datos. El conjunto de datos fue diseñado para cubrir tres tipos de clima (niebla, lluvia y (asfalto, nieve) y tres niveles de iluminación (brillante, moderada y baja), junto con tres tipos de calles relacionados con el tipo de hierba y adoquines). Las imágenes procesadas del conjunto de datos contienen tres etiquetas

(asfalto, pasto y adoquín). Las imágenes procesadas del conjunto de datos contienen tres etiquetas en su columna

relacionado con el tipo de clima, nivel de iluminación y tipo de calle. Los autores utilizaron ResNet18 y

vertebral y concluyeron que el sistema funcionó con baja precisión en el conjunto de datos necesitaban mejoras adicionales.

En [12], los autores abordan los desafíos de los vehículos autónomos durante condiciones climáticas adversas, donde los modelos perceptivos típicos luchan. Las investigaciones existentes se centran principalmente en clasificar las condiciones climáticas; sin embargo, los autores estudiaron las transiciones entre estos tipos de

clima. Propusieron un método para definir y comprender seis estados intermedios de transición climática (nublado a lluvioso, lluvioso a nublado, soleado a lluvioso, lluvioso a soleado, soleado a brumoso y de niebla a soleado). El enfoque implica interpolar datos de transición climática intermedia utilizando un codificador automático de variación, extraer características espaciales con redes convolucionales muy profundas VGG (Visual Geometry Group) y modelar la distribución temporal con una unidad recurrente cerrada para la clasificación. Los autores propusieron un nuevo conjunto de datos a gran escala llamado AIWD6 (Adverse Intermediate Weather Driving), y los resultados mostraron un modelo de transición climática eficaz.

En [13], los autores presentan un marco novedoso llamado WeatherNet, que emplea cuatro modelos CNN profundos basados en la arquitectura ResNet50. WeatherNet extrae de forma autónoma información meteorológica de la imagen de entrada y clasifica la salida en la categoría correcta. Sin embargo, el inconveniente del marco presentado es la imposibilidad de compartir funciones, ya que los cuatro modelos funcionan por separado.

Árbitro. [14] se centra en el impacto significativo de las condiciones climáticas adversas en el tráfico urbano y destaca la importancia del reconocimiento de las condiciones climáticas para aplicaciones como la asistencia AV y los sistemas de transporte inteligentes. Aprovechando los avances en el aprendizaje profundo, el artículo presenta un nuevo modelo simplificado llamado ResNet15, una versión propuesta del famoso ResNet50 [15]. El modelo propuesto tiene una capa completamente conectada que utiliza el clasificador Softmax. El artículo también presenta un nuevo conjunto de datos llamado "WeatherDataset-4" que contiene alrededor de 5000 imágenes que cubren condiciones climáticas con niebla, lluvia, nieve y sol. Aunque la red propuesta superó a la ResNet50 tradicional, el documento carece de cobertura de entornos nocturnos y arenosos.

En [16], los autores propusieron el algoritmo MCS-YOLO para mejorar la detección de objetos integrando un mecanismo de atención coordinada, una estructura multiescala para objetos pequeños y aplicando la estructura Swin Transformer [17]. A través de experimentos con el conjunto de datos BDD100K, demostraron una precisión promedio promedio (mAP) del 53,6%.

El artículo [18] es uno de los primeros artículos que aplicó CNN para la clasificación meteorológica AV. Los autores agregaron dos capas completamente conectadas para extraer características de imágenes de condiciones de servicio vial (RSC). El artículo se centró en las condiciones invernales de las carreteras, donde el problema de las carreteras nevadas se dividió en tres experimentos: (a) clasificación de dos clases, (b) clasificación de tres clases y (c) clasificación de cinco clases. El modelo superó las técnicas de clasificación tradicionales y registró una precisión del 78,5% al aplicar la clasificación de cinco clases.

En [19], YOLOv4 se ha mejorado para detectar objetos proponiendo una cabeza desacoplada y sin anclajes. El artículo utilizó BDD100k como conjunto de datos original y creó una nueva versión que se centra en tres tipos de clima (lluvioso, nevado y con niebla). Los resultados experimentales mostraron un mAP del 60,3%.

En [20], los autores extrajeron datos de movimiento de alta precisión y propusieron un nuevo mecanismo de seguimiento de vehículos llamado SORT++. Image-Adaptive YOLO (IA-YOLO) se presentó en [21] y mostró una mejora en la detección de objetos en entornos con poca luz y niebla.

Árbitro. [22] propusieron una red de subred dual (DSNet) para detectar objetos y lograron un mapa del 50,8% del tiempo con niebla. En [23] se investigó YOLOv5 para detectar objetos de varias clases, y el mAP de todas las clases obtuvo una puntuación del 25,8%. En [24], se crearon imágenes de drones y se aplicaron a una versión modificada de YOLOv5, que obtuvo un mAP de aproximadamente el 50%.

El artículo [25] comparó el rendimiento de YOLOv3, YOLOv4 y Faster R-CNN durante diferentes tipos de clima (lluvioso, con niebla, nevado). El artículo concluyó que YOLOv4 superó a YOLOv3 y Faster R-CNN.

La Tabla 1 muestra un resumen de publicaciones recientes sobre clasificación climática y detección de objetos en el entorno AV. Si bien los modelos estándar de detección de objetos se centran principalmente únicamente en el proceso de detección, nuestro trabajo y el modelo propuesto introducen varias diferencias clave en comparación con estudios relacionados recientes. Primero, hemos incorporado una nueva fase en nuestro modelo llamada "Bloque de Calidad", diseñada para evaluar y mejorar la escena observa
En segundo lugar, hemos agregado una puntuación de umbral ajustable para reducir la cantidad de imágenes que ingresan a la fase de mejora. En tercer lugar, nuestro estudio aborda de manera única las condiciones climáticas arenosas, que no han sido consideradas en publicaciones recientes.

Tabla 1. Publicaciones recientes en el entorno AV eliminan el clima arenoso de sus estudios.
Además, existe una brecha a la hora de combinar la clasificación meteorológica y la detección de objetos.

Papel	Clima Clasificación	Objeto Detección	Conducir en Nieve	Conducir en Lluvia	Conducir en Niebla	Conducir en Arena
[11]	√	×	√	√	√	×
[13]	√	×	√	√	×	×
[14]	√	×	√	√	√	×
[16]	×	√	√	√	√	×
[18]	√	×	√	×	×	×
[19]	×	√	√	√	√	×
[20]	×	√	√	√	×	×
[21]	×	√	×	×	√	×
[22]	×	√	×	×	√	×
[23]	×	√	×	√	×	×
[24]	×	√	√	√	×	×
[25]	×	√	√	√	√	×
Nuestro	√	√	√	√	√	√

3. Metodología

Nuestra metodología para desarrollar un modelo capaz de clasificar el clima y
La detección de objetos en condiciones climáticas adversas se inició mediante la aplicación de Detección en condiciones climáticas adversas.
(AMANECER) conjunto de datos [26]. Nos centramos en cubrir cuatro tipos de clima clave (arenoso, lluvioso, con niebla,
y nevado) con seis clases (peatonal, bicicleta, automóvil, motocicleta, autobús y camión). Expandir
el conjunto de datos e introducir una nueva variación de las imágenes existentes, hemos incluido datos
aumento en nuestro trabajo. Este conjunto de datos aumentado se combinó con el DAWN original.
conjunto de datos para aumentar el número de muestras de entrenamiento. Una descripción completa del aumento.
se proporcionará en la Sección 6. Luego dividimos el conjunto de datos combinado en entrenamiento y
conjuntos de validación. Nuestro porcentaje dividido es el 80% de las imágenes para entrenamiento, mientras que (20%) fueron
utilizado para validación y prueba (10% para validación y 10% para prueba). el conjunto de entrenamiento
se utilizó para entrenar tanto los modelos de clasificación climática como de detección de objetos, mientras que el
El conjunto de validación cumplió el papel fundamental de prevenir el sobreajuste. Después de eso, pasos de optimización.
están involucrados para encontrar el mejor rendimiento del modelo cambiando los hiperparámetros.
Electrónica 2024, 13, x PARA REVISIÓN POR PARES Por último, evaluamos los modelos optimizados utilizando la precisión media estándar. 7 de 21
(mAP), precisión y métricas de recuperación. La Figura 4 muestra la secuencia de nuestra metodología.

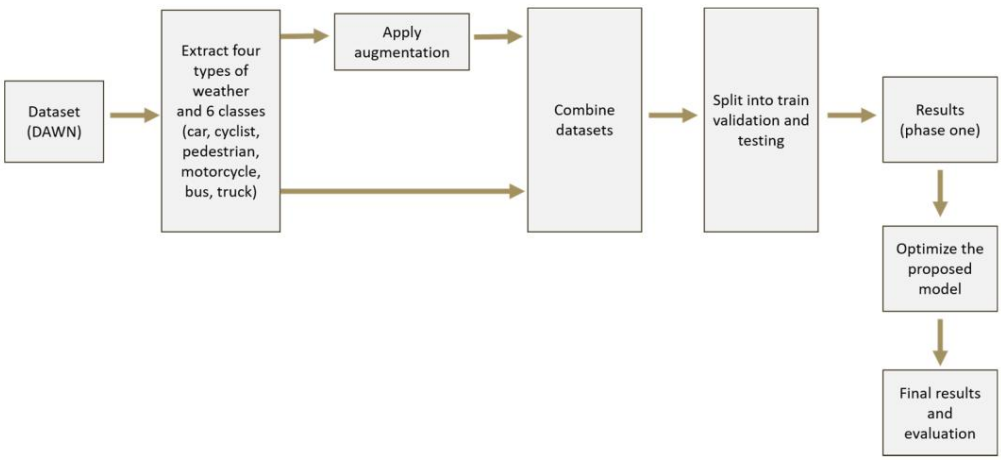


Figura 4. Secuencia de nuestra metodología en este artículo.
Figura 4. Secuencia de nuestra metodología en este artículo.

Para optimizar la eficiencia computacional dados los recursos limitados de la GPU, empleamos Para optimizar la
eficiencia computacional dados los recursos limitados de la GPU, empleamos
La plataforma Colab basada en la nube de Google como nuestro entorno experimental. Colab proporciona una
marco de aprendizaje automático PyTorch y GPU de alto rendimiento (como el
T4). A través de Colab pudimos ejecutar nuestros experimentos de manera efectiva, específicamente
A través de Colab pudimos ejecutar nuestros experimentos de manera efectiva, específicamente con el
con la integración con CUDA (Compute Unified Device Architecture), que ayuda a la integración con CUDA (Compute
Unified Device Architecture), que ayuda a acelerar
en acelerar el proceso computacional de nuestra canalización y la parte de CNN mientras detecta
objetos (específicamente dentro de tareas como convolución, agrupación, normalización y capas
de activación).

Hay varias métricas disponibles para cuantificar la eficacia de los modelos de detección de
objetos. En nuestro artículo, priorizamos tres métricas principales: (a) precisión promedio media (mAP),
(b) precisión y (c) recuperación. mAP se erige como una métrica de evaluación predominante dentro de la do-

el proceso computacional de nuestra canalización y la parte CNN mientras detecta objetos (específicamente dentro de tareas como capas de convolución, agrupación, normalización y activación).

Hay varias métricas disponibles para cuantificar la eficacia de los modelos de detección de objetos. En En nuestro artículo, priorizamos tres métricas principales: (a) precisión promedio media (mAP), (b) precisión y (c) recuperación. mAP se erige como una métrica de evaluación predominante dentro del dominio del objeto Detección, que ofrece una evaluación holística de la competencia del modelo en la identificación de objetos. y localización. mAP combina precisión y recuperación calculando la precisión promedio (AP) para cada clase o categoría de objeto, derivando posteriormente la media en todas las clases. AP sirve como medida de la calidad de la detección, encapsulando tanto la precisión de la precisión objetos identificados y la integridad de la detección en la escena. A través del cálculo de mAP, el desempeño de nuestro oleoducto se puede comparar y evaluar numéricamente en diversos dominios y escenarios.

También consideramos la precisión y el recuerdo como métricas indispensables en el contexto de detección de objetos. La precisión es la proporción o el porcentaje de elementos recuperados que son relevantes para la clase correcta, mientras que el recuerdo mide el porcentaje de objetos relevantes que son recuperado con éxito. La precisión se expresa como la relación entre los verdaderos positivos (TP) y la suma de verdaderos positivos y falsos positivos (FP), representados como:

$$\text{Precisión} = TP/(TP + FP)$$

La recuperación es la relación entre TP y la suma de verdaderos positivos y falsos negativos (FN), representada como:

$$\text{Recuperar} = TP/(TP + FP)$$

4. Conjunto de datos

Para el conjunto de datos, como mencionamos anteriormente, utilizamos DAWN en nuestro desarrollo y experimentación. El conjunto de datos DAWN cubre cuatro tipos de clima adverso: tormenta de arena, lluvia, nieve y niebla. La Figura 5 muestra una muestra de los diversos tipos de clima cubiertos en AMANECER. El conjunto de datos contiene 1027 imágenes que cubren los cuatro tipos de clima y diferentes contextos ambientales como carreteras, autopistas y paisajes urbanos, asegurando una Amplia representación de escenarios del mundo real.

Figura 5. El conjunto de datos DAWN proporciona diversas condiciones climáticas adversas, como niebla, lluvia, nieve y arena [26].

Figura 5. El conjunto de datos DAWN proporciona diversas condiciones climáticas adversas, como niebla, lluvia, nieve y arena [26].

Aunque muchos otros conjuntos de datos cubren condiciones climáticas adversas, el conjunto de datos DAWN tiene la ventaja de incluir imágenes de tormentas de arena o de clima arenoso, que a menudo están ausentes de otros conjuntos de datos. Esta característica única de DAWN ha permitido que nuestro modelo aborde la clasificación climática y la detección de objetos en múltiples tipos de entornos geográficos.

El conjunto de datos DAWN utilizado originalmente consta de 1027 imágenes con un tamaño de 640 × 640.

La anotación de imagen contiene la clase del objeto y los límites correspondientes de: x, y, ancho y alto del cuadro delimitador (x_center, y_center, ancho, alto). Higo-

Figura 5. El conjunto de datos DAWN proporciona diversas condiciones climáticas adversas, como niebla, lluvia, nieve y arena [26].

Aunque muchos otros conjuntos de datos cubren condiciones climáticas adversas, el conjunto de datos DAWN tiene la ventaja de incluir tormentas de arena o imágenes del clima arenoso, que a menudo están ausentes en otros conjuntos de datos. Esta característica única de DAWN ha permitido que nuestro modelo aborde meteorológica y detección de objetos en múltiples tipos de entornos geográficos.

El conjunto de datos DAWN utilizado originalmente consta de 1027 imágenes con un tamaño de 640 × 640. La anotación de imagen contiene la clase del objeto y los límites correspondientes. La anotación de imagen contiene la clase del objeto y los límites correspondientes de: x, y, ancho y alto del cuadro delimitador (x_center, y_center, ancho, alto). Fig: x, y, ancho y alto del cuadro delimitador (x_center, y_center, ancho, alto). Figura 6

La figura 6 representa una muestra de nuestras imágenes etiquetadas, consideradas como una referencia de verdad fundamental. representa una muestra de nuestras imágenes etiquetadas, consideradas como una referencia de verdad sobre el terreno.

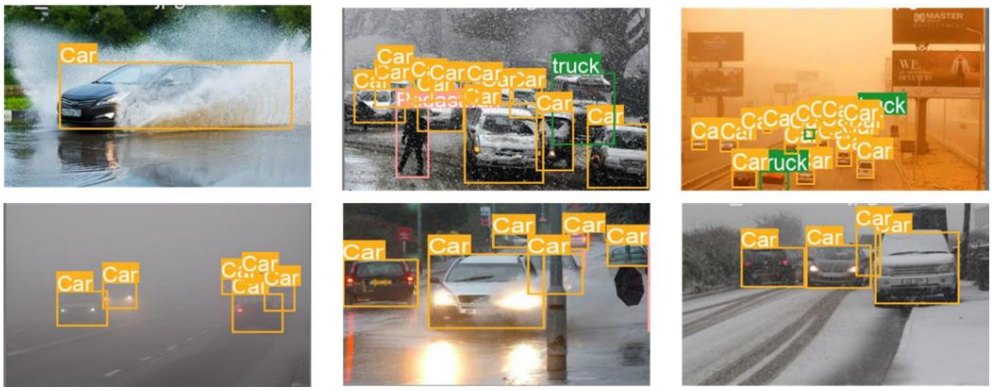


Figura 6. Muestra de imágenes etiquetadas.
Figura 6. Muestra de imágenes etiquetadas.

5. Modelo propuesto

Nuestro oleoducto propuesto establece una clasificación climática integral y un objeto integración de cuatro tareas principales: (1) evaluación de la calidad de la imagen, (2) detección de mejora de la imagen- ment, (3) clasificación climática y (4) detección de objetos. Hemos combinado la tarea 1 y 2 en un bloque llamado "Bloque de calidad", mientras que las tareas 3 y 4 se combinan en otro llamado "Bloque de clasificación y detección". Una vez que la imagen ingresa al modelo, se realizará una evaluación. Se realizarán pruebas para evaluar la calidad de la imagen. El propósito de este paso es tomar una decisión si la imagen capturada necesita una mejora o no. Se ha empleado el método del Evaluador de calidad espacial de imagen de referencia ciega (BRISQUE) [27] para determinar la calidad de la imagen sin cuantificación . Si la imagen ingresada tiene una puntuación superior a un punto umbral (baja calidad), la imagen será rechazada y transferida a una etapa de mejora; en caso contrario, será aprobado y transferido directamente a la clasificación y detección. bloquear. El punto de umbral se puede cambiar y modificar según la situación de la escena; para Por ejemplo, en nuestro experimento hemos utilizado un umbral de 42,7, como explicaremos en la Sección 7. "Experimentos y resultados". Vale la pena señalar que, en el método BRISQUE, generalmente un Una puntuación más baja indica una mejor calidad perceptiva, mientras que una puntuación BRISQUE más alta indica peor calidad perceptiva. El algoritmo BRISQUE tiene varias ventajas que lo convirtieron en un solución adecuada para nuestro modelo y para la evaluación de escenas climáticas adversas. En primer lugar, es un métrica de calidad de imagen sin referencia, ya que no requiere una imagen de referencia perfecta para comparación. Esto es muy ventajoso en condiciones climáticas adversas, donde obtener Obtener imágenes ideales y sin distorsiones puede ser un desafío, si no imposible. BRISQUE funciona mediante analizar las estadísticas de escena natural (NSS) de una imagen y compararlas con las estadísticas esperadas de imágenes naturales (sin distorsiones). Cualquier desviación de esta naturalidad es marcados como indicadores de degradación de la calidad, lo que los convierte en una buena opción para detectar los tipos de distorsiones introducidas por los fenómenos meteorológicos. Además, BRISQUE ofrece herramientas computacionales. eficiencia en comparación con varias otras opciones, que pueden ser importantes al trabajar con grandes conjuntos de datos de imágenes o en escenarios donde se desea una evaluación de la calidad en tiempo real. Para En la fase de mejora de la imagen, hemos utilizado el Adversario Generativo de Súper Resolución. Técnica de red (SRGAN) [28], que consta de redes generadoras y discriminadoras.

La red generadora tiene como objetivo mejorar las imágenes de baja resolución, mientras que el discriminador La red tiene como objetivo refinar la salida del generador, lo que resulta en una claridad de imagen mejorada. Después del Bloque de Calidad, la imagen es procesada por dos redes YOLOv5. Un Yolo La red, entrenada extensamente en un conjunto de datos de imágenes etiquetadas con el clima, clasifica con precisión condiciones climáticas como arena, lluvia, nieve o niebla. Al mismo tiempo, un separado

técnica [28], que consta de redes generadoras y discriminadoras. La red generadora tiene como objetivo mejorar las imágenes de baja resolución, mientras que la red discriminadora tiene como objetivo refinar la salida del generador, lo que resulta en una claridad de imagen mejorada. Después del Bloque de Calidad, la imagen es procesada por dos redes YOLOv5. Una red YOLO,

capacitado ampliamente en un conjunto de datos de imágenes etiquetadas con el clima, clasifica con precisión las condiciones climáticas como arena, lluvia, nieve o niebla. Simultáneamente, una red YOLO separada, entrenada para identificar y localizar objetos con cuadros delimitadores, detectará la red YOLO objetivo, entrenada para identificar y localizar objetos con cuadros delimitadores, detectará

objetos, como automóviles, ciclistas, peatones, motocicletas, autobuses y camiones. Generalmente, nuestros objetivos son objetivos, como automóviles, ciclistas, peatones, motocicletas, autobuses y camiones.

El modelo propuesto ofrece un enfoque doble, priorizando la calidad de la imagen. En general, nuestro modelo propuesto ofrece un enfoque doble, priorizando la calidad de la imagen.

transición sin problemas a una clasificación meteorológica y detección de objetos sólidas basadas en YOLO antes de realizar una transición sin problemas a una clasificación meteorológica y detección de objetos sólidas basadas en YOLO

ción para un análisis de imágenes confiable. La Figura 7 muestra una ilustración de nuestra propuesta.

detección para un análisis de imágenes confiable. La Figura 7 muestra una ilustración de nuestra propuesta.

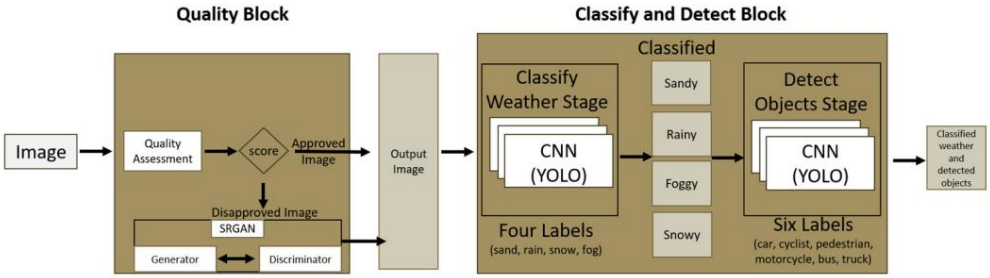


Figura 7. Nuestro modelo propuesto.

Figura 7. Nuestro modelo propuesto.

6. AMANECER aumentado

6. DAWN aumentado Dado

el numero limitado de imágenes meteorológicas adversas en el conjunto de datos DAWN, construimos nueva versión de DAWN para ampliar nuestro conjunto de datos experimental. Datos

Dada la cantidad limitada de imágenes climáticas adversas en el conjunto de datos de DAWN, utilizamos una aumento es un método ampliamente utilizado para ampliar artificialmente conjuntos de datos mediante la creación de nuevos

Construimos una nueva versión de DAWN usando aumento para ampliar nuestro conjunto de datos experimental. El

El aumento de datos es un método ampliamente utilizado para ampliar artificialmente conjuntos de datos mediante la

como [29,30], la mejora incremental de YOLOv5 [8] han empleado el aumento de datos ya sea para su nuevas imágenes de entrenamiento de un conjunto de datos actualmente disponible. Varios artículos, y la mejora incremental de YOLOv3 [8] han empleado conjuntos de datos de clasificación meteorológica de aumento de

datos o detección de objetos. Nuestra versión aumentada ha aumentado

variantes de conjuntos de datos DAWN de 1027 imágenes a 2046 imágenes, lo que supone un aumento de casi 21

sion ha aumentado la cantidad de conjuntos de datos DAWN de 1027 imágenes a 2046 imágenes, antes y después de

nuestro aumento aplicado

nuestro aumento aplicado de casi el doble del tamaño actual. La Figura 8 muestra una descripción general del conjunto de datos DAWN antes y después de nuestro aumento aplicado.

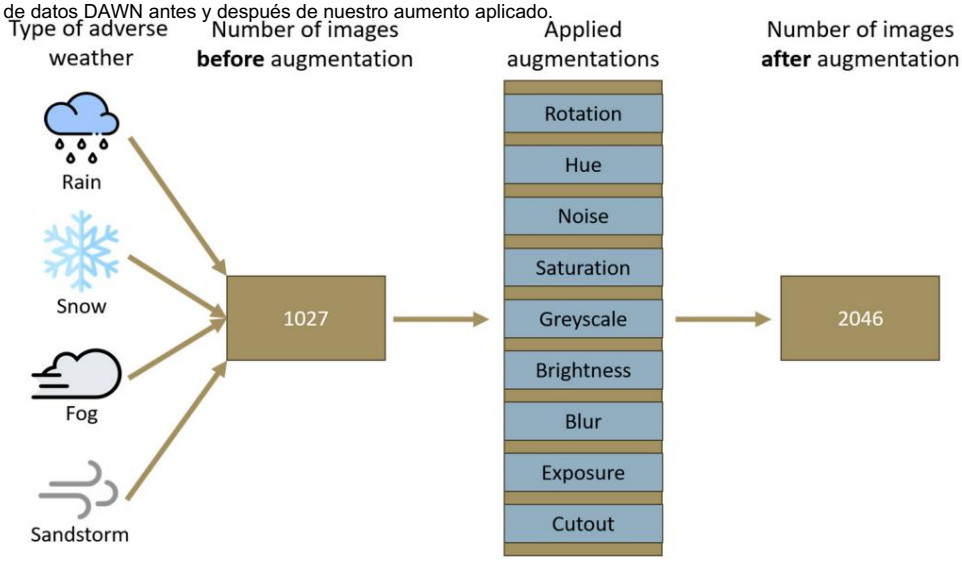


Figura 8. El conjunto de datos DAWN aumentó de 1027 a 2046 imágenes.

Figura 8. El conjunto de datos DAWN aumentó de 1027 a 2046 imágenes.

Las técnicas de aumento abarcan una variedad de ajustes de imagen, como la imagen

Las técnicas de aumento abarcan una variedad de ajustes de imagen, como escalado, rotación, recorte, volteo, ajustes de color, ruido o desenfoque, y muchos otros

ajustes de color, ruido o desenfoque, y muchos otros escalados, rotación, recorte, volteo, ajuste de color, ruido o operaciones. La aplicación de aumento para la clasificación meteorológica y la detección de objetos puede ser

operaciones. La aplicación de aumento para la clasificación del clima y la detección de objetos puede ser muy ventajoso por las siguientes motivos:

la diversidad y variabilidad de los datos de entrenamiento, ayudando a la generalización del modelo • Mejorar

• Aumentar la resiliencia del modelo frente a factores que afectan la apariencia del objeto, como condiciones de

adaptación a escenarios no representados.

• Impulsar la resiliencia del modelo contra factores que afectan la apariencia del objeto, como iluminación variables, oclusiones o cambios de punto de vista.

condiciones de iluminación variables, oclusiones o cambios de punto de vista.

• Abordar el desequilibrio de clases sobremuestreando las clases minoritarias o submuestreando los mayoritarios.

• Mitigar el sobreajuste mediante la introducción de regularización y ruido en los datos de entrenamiento.

Las siguientes secciones describen los aumentos que realizamos en este artículo.

- Abordar el desequilibrio de clases sobremuestreando las clases minoritarias o submuestreando los mayoritarios.
- Mitigar el sobreajuste mediante la introducción de regularización y ruido en los datos de entrenamiento.

Las siguientes secciones describen los aumentos que realizamos en este artículo.

6.1. Rotación

Esto se utiliza para presentar el objeto desde diferentes ángulos de visión. En escenarios reales, los objetos pueden aparecer en diferentes ángulos o rotaciones, y agregar esto a nuestro aumento puede ayudar al modelo a manejar mejor estas variaciones de vista.

6.2. Matiz

Hue es una técnica de aumento de imágenes basada en colores que altera el tono o tono de color de una imagen preservando su brillo y saturación.

6.3. Ruido

También hemos incorporado ruido sintético en nuestro proceso de aumento para ampliar nuestro conjunto de datos. Este tipo de aumento mejora la resistencia de nuestro modelo al ruido y mejora su capacidad para adaptarse a nuevos datos o escenarios.

6.4. Saturación

La saturación ajusta la intensidad de los colores dentro de una imagen. Al saturar una imagen, escalamos efectivamente los valores de píxeles mediante un factor aleatorio dentro de un rango específico. Aumentar el valor de saturación de una imagen puede hacer que los colores sean más vibrantes y vívidos, mientras que disminuirlo puede hacer que los colores sean más apagados y apagados. Aumentamos la saturación de nuestro conjunto de datos en aproximadamente un 25%.

6.5. Escala de grises

Incorporamos aumento de escala de grises, que convierte una imagen en escala de grises. Esta técnica se utiliza comúnmente para aumentar el contraste de una imagen y mejorar sus detalles.

6.6. Brillo Al

aumentar aleatoriamente el brillo de las imágenes, sometimos nuestro modelo a una gama más amplia de condiciones de iluminación, mejorando así su resistencia a los cambios de iluminación.
Aumentamos las imágenes, haciéndolas aproximadamente un 15% más brillantes.

6.7. Difuminar

El desenfoque se utiliza para introducir efectos desenfocados en las imágenes. Para nuestros datos aumentados, nosotros Usé desenfoque gaussiano con hasta 1,25 px.

6.8. Exposición

Además, modificamos artificialmente el nivel de exposición de las imágenes, configurándolo en el rango del 10% al −10%.

6.9. Separar

También hemos cortado pequeñas partes de objetos de la escena. El propósito de esto es agregar oclusión a nuestro experimento, que consiste en bloquear, cubrir u oscurecer un objeto desde la vista de la cámara.

La Tabla 2 muestra nuestros valores de configuración de aumento y sus impactos en las imágenes.

Tabla 2. Resumen de aumentos aplicados y su impacto en la imagen.

Aumento	Valor	Impacto
Rotación	90 grados	Ayuda al modelo a ser insensible a la orientación de la cámara.
Matiz	15%	Ajuste aleatorio de colores.
Ruido	Ruido aleatorio	Más obstáculos añadidos a la imagen.

Tabla 2. Cont.

Aumento	Valor	Impacto
Saturación	25%	Cambia la intensidad de los píxeles.
Escala de grises	15%	Convierte la imagen a un solo canal
Brillo	15%	La imagen aparece más clara
Difuminar	1.25px	Promedia los valores de píxeles con los vecinos
Exposición	10%	Resistente a los cambios de iluminación y configuración de la cámara.
Separar	Cortar partes aleatorias de la imagen.	Más resistente para detectar medios objetos.

7. Experimentos y resultados

Para probar nuestro modelo, hemos realizado varios experimentos, comenzando por establecer nuestro Puntuación umbral de BRISQUE a 42,75. Esta puntuación es la puntuación de calidad promedio de DAWN. conjunto de datos, y cualquier imagen por encima de esta puntuación promedio pasará por la etapa de mejora. La Tabla 3 explica la razón detrás de elegir 42,75 como nuestro punto umbral e ilustra la impacto de la calidad de la imagen midiendo las puntuaciones de BRISQUE antes y después del aumento. El La tabla compara imágenes del conjunto de datos DAWN (tormentas de arena, lluvia, nieve y niebla) con nuestra Imágenes aumentadas extendidas del mismo conjunto de datos con el objetivo de simular el clima adverso. condiciones. En todas las condiciones climáticas, las imágenes aumentadas generalmente muestran niveles más altos. Puntuaciones BRISQUE, lo que indica una disminución en la calidad de la imagen en comparación con el DAWN original. imágenes. Como muestra la tabla, las imágenes aumentadas son peores en aproximadamente un 9% cuando se trata de la calidad media de la escena. Esta baja calidad de las imágenes aumentadas se puede atribuir a los aumentos realizados (desenfoque, tono, saturación, ruido, corte, brillo y exposición), que son efectos habituales durante condiciones climáticas adversas. Las diferencias observadas subrayan la importancia de diseñar un modelo de evaluación de la calidad para preservar la calidad de la imagen, particularmente en condiciones climáticas adversas donde la claridad visual es crucial para una observación precisa de la escena y detección de objetos.

Tabla 3. Comparación de la calidad de imagen con y sin aumento. Los resultados muestran una calidad de imagen promedio de 46,59 para el conjunto de datos DAWN aumentado en comparación con 42,75 para el original Conjunto de datos del amanecer.

	Arenoso	Neblinoso	Nevado	Lluvioso	Promedio
Imágenes del amanecer Nivel de calidad	44.05	45.21	40.18	41,57	42,75
Imágenes aumentadas de DAWN Nivel de calidad	48,71	49,83	43.19	44,64	46,59

El escenario experimental para el conjunto de datos DAWN aumentado se ejecutó dentro del Entorno de Google Colab, aprovechando la potencia computacional de una GPU Tesla T4. Nosotros han realizado varias modificaciones a la arquitectura YOLOv5, con el objetivo de crear una adecuada modelo para nuestro dominio. Esta modificación incluye el cambio de las funciones de activación y Pruebe el modelo con las funciones SiLU y LeakyRelu. También hemos modificado la columna vertebral. y diríjase a probar el rendimiento de las arquitecturas BottleneckCSP y C3. Además Además, los hiperparámetros como las épocas y el tamaño del lote se han cambiado a lo largo nuestros experimentos.

Después de diseñar nuestro modelo, iniciamos nuestra fase experimental implementando BottleneckCSP como arquitectura principal y troncal. Nuestro modelo demostró resultados prometedores, logrando una precisión media media (mAP) del 55,6% y del 45,6% cuando entrenado para 128 épocas con un tamaño de lote de 32, utilizando la activación SiLU y LeakyReLU funciones, respectivamente. En la Tabla 4, se puede ver claramente que, cuando implementamos BottleneckCSP en nuestro modelo, el mAP estaba aumentando para las funciones SiLU y LeakyRelu cada vez que aumentamos el número de épocas. También se puede ver que LeakyRelu tiene menor rendimiento que SiLU con la columna vertebral y el cabezal BottleneckCSP. La tabla 4 muestra la Resultados completos de nuestro modelo utilizando BottleneckCSP.

Tabla 4. Rendimiento de nuestro modelo utilizando BottleneckCSP como columna vertebral y cabeza.

Función de activación de la columna vertebral y la cabeza		Época	Lote	mapa
Cuello de botellaCSP	SiLU	32	dieciséis	33,7%
Cuello de botellaCSP	SiLU	32	32	34,5%
Cuello de botellaCSP	SiLU	64	dieciséis	40,2%
Cuello de botellaCSP	SiLU	64	32	43,9%
Cuello de botellaCSP	SiLU	128	dieciséis	55,2%
Cuello de botellaCSP	SiLU	128	32	55,6%
Cuello de botellaCSP	LeakyRelu	32	dieciséis	24,0%
Cuello de botellaCSP	LeakyRelu	32	32	25,1%
Cuello de botellaCSP	LeakyRelu	64	dieciséis	34,7%
Cuello de botellaCSP	LeakyRelu	64	32	34,9%
Cuello de botellaCSP	LeakyRelu	128	dieciséis	38,7%
Cuello de botellaCSP	LeakyRelu	128	32	45,6%

Continuamos nuestros experimentos pasando a incluir Concentrado-Completo Convolución (C3) [31] como columna vertebral y cabeza en nuestro modelo propuesto. El modelo logró un mejor resultado, logrando un 71,8% de mAP usando SiLU con solo 32 épocas y 16 lotes, como La tabla 5 lo muestra. Esta puntuación es superior a la de LeakyRelu en 7,4 puntos porcentuales, con el mismo métrica. Este resultado también es más alto que el puntaje más alto logrado con BottleneckCSP. columna vertebral (Cuadro 4), que fue del 55,6%. Seguimos aumentando el número de épocas.

y lotes hasta que logramos el 74,6% después de 64 épocas con 16 lotes, que es el más alto puntuación mAP en este documento. Como se demostrará a continuación, esta puntuación mAP es la más alta en comparación con otras publicaciones recientes sobre detección de objetos que utilizaron DAWN como conjunto de datos base. El mapa, la precisión y la recuperación de nuestro modelo se muestran en la Figura 9. El gráfico superior izquierdo muestra el resultado de precisión, con una puntuación del 85 %, mientras que el gráfico superior derecho muestra la recuperación que alcanzó el 68 %. El gráfico inferior muestra el mAP, que alcanzó el 74,6%. La Figura 10a muestra nuestra puntuación F1 y podemos ver claramente que la mayoría de las clases se producen en umbrales de confianza entre 0,4 y 0,6. Esto sugiere que establecer el umbral de confianza del modelo dentro de este rango probablemente produciría el mejor equilibrio entre precisión y recuperación. También muestra que el modelo destaca en la detección de automóviles, mientras que el rendimiento general sigue un patrón similar al de los camiones, y el promedio en todas las clases. La Figura 10b representa nuestro mapa, con intersección sobre unión de 0,5 a 0,95.

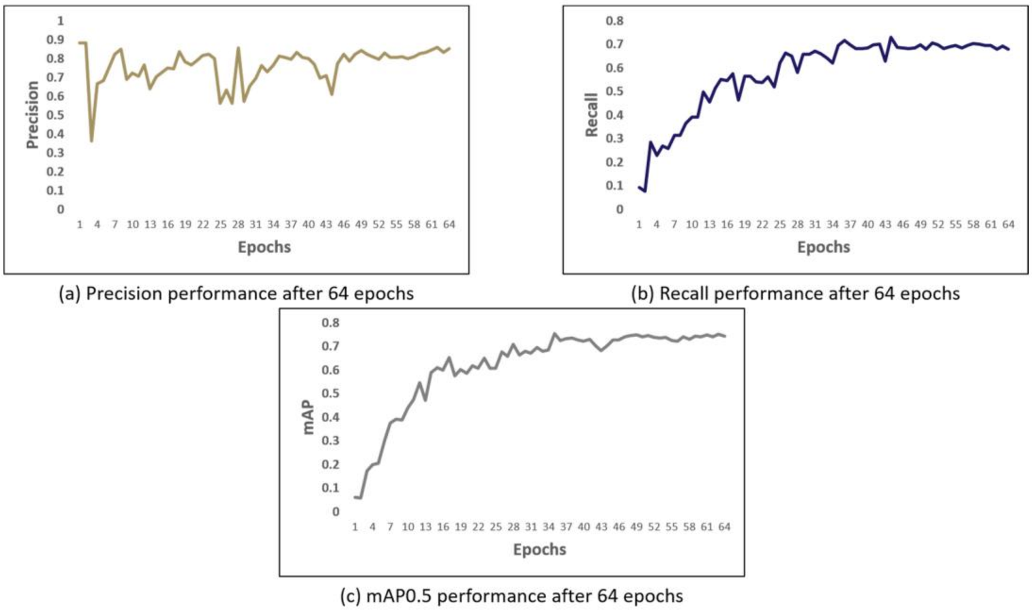


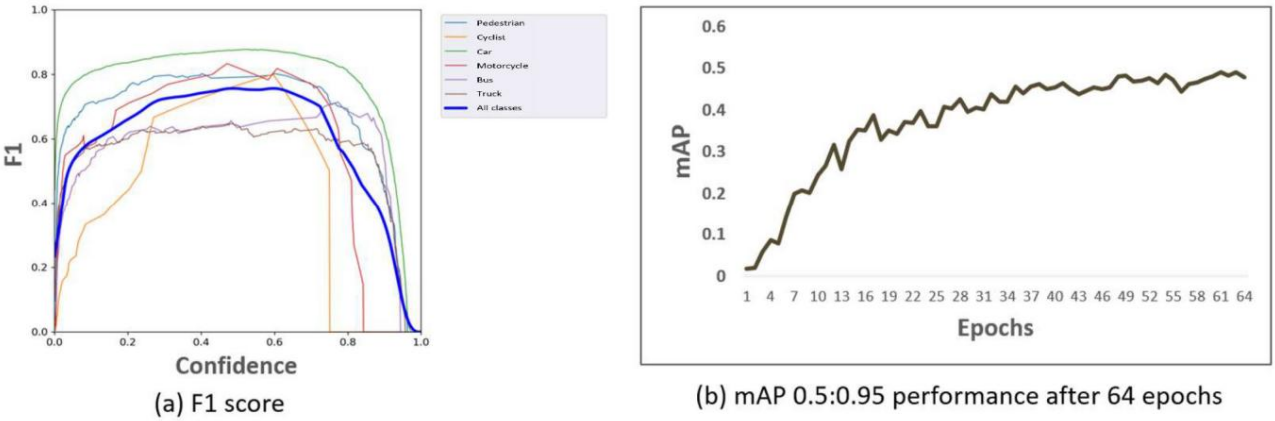
Figura 9. Precisión, recuperación y mAP a 0,5 de nuestro modelo después de 64 épocas.

Tabla 5. Rendimiento de nuestro modelo utilizando C3 como columna vertebral y cabeza.

Columna vertebral y cabeza	Función de activación	Época	Lote	mapa
C3	SiLU	32	dieciséis	71,8%
C3	SiLU	32	32	68,6%

Tabla 5. Rendimiento de nuestro modelo utilizando C3 como columna vertebral y cabeza.

Función de activación de la columna vertebral y la cabeza		Época	Lote	mapa
C3	SiLU	32	diecisés	71,8%
C3	SiLU	32	32	68,6%
C3	SiLU	64	diecisés	74,6%
C3	SiLU	64	32	74,1%
C3	SiLU	128	diecisés	74,0%
C3	SiLU	128	32	73,1%
C3	LeakyRelu	32	diecisés	64,4%
C3	LeakyRelu	32	32	67,1%
C3	LeakyRelu	64	diecisés	62,9%
C3	LeakyRelu	64	32	63,2%
C3	LeakyRelu	128	diecisés	21 72,9%
C3	LeakyRelu	128	32	72,4%



(a) F1 score (b) mAP 0.5:0.95 performance after 64 epochs

Figura 10. Puntuación F1 y mAP en 0,5:0,95 de nuestro modelo después de 64 épocas.

Figura 10. Puntuación F1 y mAP en 0,5:0,95 de nuestro modelo después de 64 épocas.

Las tablas mencionadas anteriormente, Tablas 4 y 5, resaltan una observación notable: aumentar el número de épocas no se correlaciona consistentemente con un promedio más alto precisión (mAP). Sorprendentemente, nuestro modelo logró su puntuación mAP más alta cuando se entrenó durante 64 épocas en lugar de 128 épocas, contrariamente a la expectativa inicial. Este descubrimiento subraya la importancia de un cuidadoso ajuste de hiperparámetros y experimentación de validación para identificar el régimen de entrenamiento más efectivo para una arquitectura de modelo y un conjunto de datos determinados.

Las tablas mencionadas anteriormente, Tablas 4 y 5, resaltan una observación notable: aumentar el número de épocas no se correlaciona consistentemente con un promedio más alto precisión (mAP). Sorprendentemente, nuestro modelo logró su puntuación mAP más alta cuando se entrenó durante 64 épocas en lugar de 128 épocas, contrariamente a la expectativa inicial. Este descubrimiento destaca la importancia de un cuidadoso ajuste de hiperparámetros y experimentación de validación para identificar el régimen de entrenamiento más efectivo para una arquitectura de modelo y un conjunto de datos determinados.

En publicaciones recientes que utilizan el conjunto de datos DAWN, la precisión promedio promedio (mAP) de nuestro método del 74,6% es la más alta alcanzada, como se detalla en la Tabla 6.

En publicaciones recientes que utilizan el conjunto de datos DAWN, el promedio medio (mAP) de nuestro método del 74,6% es el más alto alcanzado, como se detalla en la Tabla 6.

Tabla 6. Comparación de nuestro resultado con algunas publicaciones recientes que utilizaron el conjunto de datos DAWN.

Tabla 6. Comparación de nuestro resultado con algunas publicaciones recientes que utilizaron el conjunto de datos DAWN.

Árbitro. mapa ref.	mapa	Objetivo del conjunto de datos	Conjunto de datos	Apuntar
[32]	32,75%	AMANECER	32,75%	Enfoque conjunto para mejorar la detección de objetos en vehículos autónomos en condiciones climáticas adversas DAWN
[33]	Niebla 29,66% Lluvia 41,21% Nieve 43,01% Arena 24,13%	AMANECER	Niebla 29,66% Lluvia 41,21% Nieve 43,01% Arena 24,13%	Modificar YOLO y usar varios conjuntos de datos para detectar objetos en el entorno AV.
[34]	39,19%	AMANECER	39,19%	Arquitectura para la construcción de conjuntos de datos utilizando GAN y CycleGAN.
[35]	55,85%	AMANECER	55,85%	Transformador de detección de poca luz (LDETR) para mejorar el rendimiento de detección.
[36]	72,8%	AMANECER	72,8%	Mejora de YOLO mediante algoritmos metaheurísticos.
Nuestro	74,6%	AMANECER	Nuestro 74,6%	Modificando YOLO usando el conjunto de datos DAWN para clasificar el clima y detectar objetos en el entorno audiovisual.

Modificando YOLO usando el conjunto de datos DAWN para clasificar el clima y detectar objetos en el entorno AV.

Para la evaluación de la clasificación climática, el modelo propuesto obtuvo una precisión del 74,3% después de 64 épocas, como muestra la Figura 11. La modelo ha clasificado con éxito la mayoría de las escenas; sin embargo, hay algunos casos en los que el modelo no logró clasificar el tiempo real. Por ejemplo, si nos fijamos en la Tabla 7, que muestra el resultado experimental de la clasificación del tiempo, en la imagen número 5 el tiempo real fue una fuerte tormenta de arena, mientras que el modelo clasificó

Para la evaluación de la clasificación climática, el modelo propuesto obtuvo una precisión del 74,3%. después de 64 épocas, como muestra la Figura 11 . La modelo ha clasificado con éxito la mayoría de las escenas; sin embargo, hay algunos casos en los que el modelo no logró clasificar el tiempo real. Para Por ejemplo, si miramos la Tabla 7, que muestra el resultado experimental de la clasificación del clima, en la imagen número 5 el clima real fue una fuerte tormenta de arena, mientras que el modelo clasificó es como un clima brumoso. Este caso es un ejemplo de donde el brillo y la iluminación del La escena podría ser un desafío para los modelos de clasificación climática en condiciones climáticas adversas.

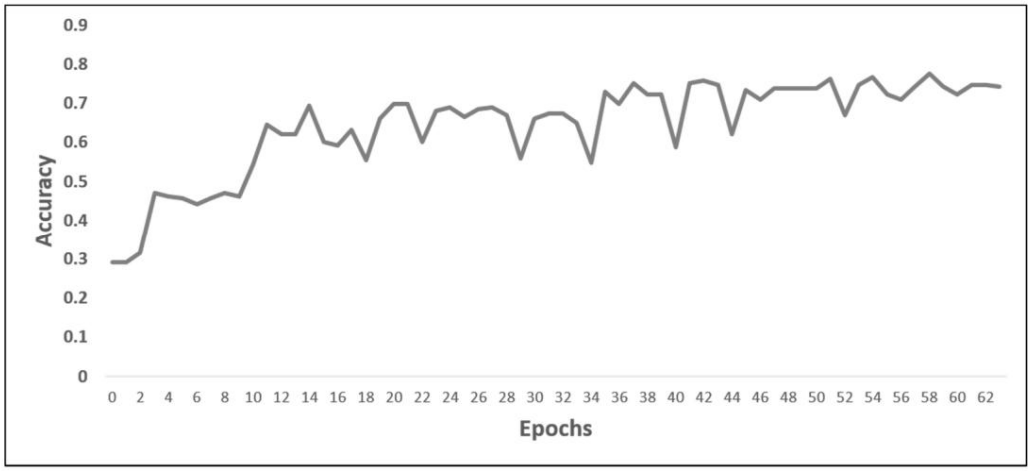


Figura 11. Para la clasificación meteorológica, nuestro modelo obtuvo una puntuación del 74,3% después de 64 épocas.
Figura 11. Para la clasificación meteorológica, nuestro modelo obtuvo una puntuación del 74,3% después de 64 épocas.
Figura 11. Para la clasificación meteorológica, nuestro modelo obtuvo una puntuación del 74,3% después de 64 épocas.
Figura 11. Para la clasificación meteorológica, nuestro modelo obtuvo una puntuación del 74,3% después de 64 épocas.

Tabla 7. Resultados de clasificación.
Tabla 7. Resultados de clasificación.
Tabla 7. Resultados de clasificación.
Tabla 7. Resultados de clasificación.

Número Imagen	Número Imagen	Suelo Clasificado
1		Clima arenoso
2		Clima con niebla
3		Clima lluvioso

Electrónica 2024, 13, x PARA REVISIÓN POR PARES 16 de 21

POR PARES Tabla 7. Cont.

Número	Imagen	Suelo	Clasificado
4		Clima nevado	Clima nevado
5		Tiempo arenoso	Tiempo brumoso

Las Figuras 12 a 14 son ejemplos de nuestros experimentos en los que el modelo clasificó con éxito el tipo de clima en la escena y detectó los objetos. La esquina superior izquierda muestra los puntajes de probabilidad para el clima actual, clasificándolos de mayor a menor probabilidad. Por ejemplo, en la Figura 12, el modelo clasificó la escena como clima arenoso en un 87%, lo cual es correcto y coincide con la verdad del terreno. También se detecta con éxito que los dos vehículos en la escena son correctos y coinciden con la verdad del terreno. También se detecta con éxito a los dos vehículos que aparecen en escena, con porcentajes de detección del 96% y 94%, respectivamente, apareciendo en escena, con porcentajes de detección del 96% y 94%, respectivamente.

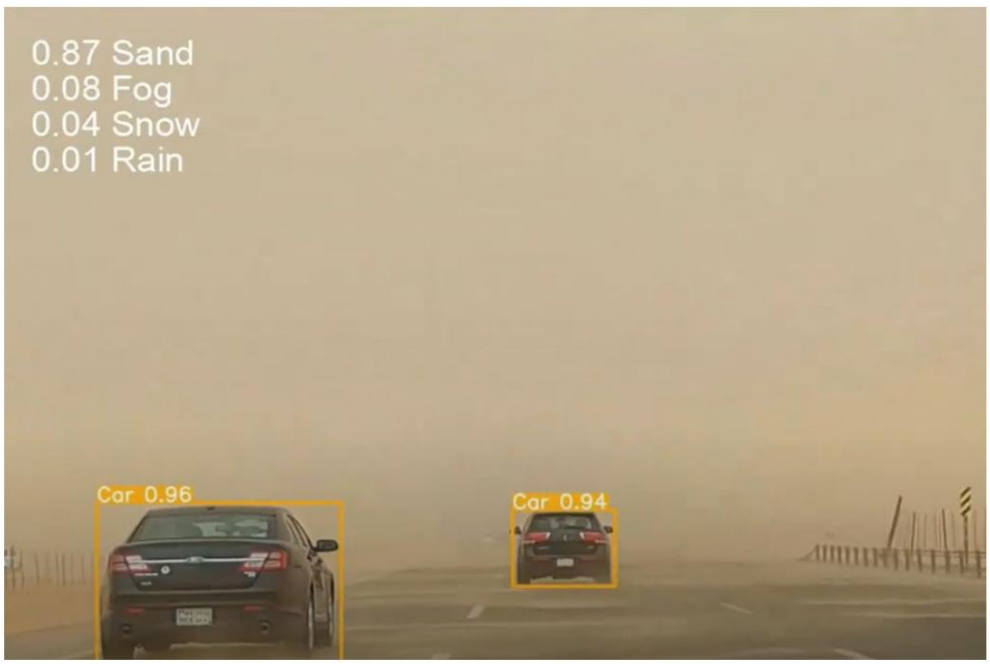


Figura 12. El modelo clasificó con éxito el clima arenoso y detectó objetos en la escena.

Figura 12. El modelo clasificó con éxito el clima arenoso y detectó objetos en la escena.

Figura 13. El modelo clasificó con éxito el clima con niebla y detectó objetos en la escena.
Figura 13. El modelo clasificó con éxito el clima con niebla y detectó objetos en la escena.

Figura 14. El modelo clasificó con éxito el clima lluvioso y detectó objetos en la escena.
Figura 14. El modelo clasificó con éxito el clima lluvioso y detectó objetos en la escena.
Figura 14. El modelo clasificó con éxito el clima lluvioso y detectó objetos en la escena.

8. Discusión

Si bien la sección anterior demuestra el potencial de nuestro método para detectar varios vehículos en condiciones climáticas adversas, en los siguientes puntos discutiremos ideas clave y Observaciones que surgieron durante el desarrollo de este trabajo:

• Si observamos nuestra puntuación de F1 (Figura 10), la clase “automóvil” logra consistentemente las puntuaciones más altas de F1 en diferentes niveles de confianza, lo que indica que el modelo es particularmente hábil para detectar autos con precisión. Por el contrario, la clase “camión” generalmente exhibe las puntuaciones más bajas de F1, lo que sugiere que el modelo podría tener más dificultades para distinguir los camiones, o enfrentar más falsos positivos/negativos en esta categoría. La curva de “todas las clases” representa el rendimiento promedio en todas las clases de objetos, lo que demuestra una tendencia similar a la de las clases individuales, con la puntuación máxima de F1 alrededor del umbral de confianza de 0,6.

• Como muestra la Tabla 6 , nuestro trabajo propuesto logró un mAP del 74,6%. Este resultado supera el rendimiento de otras publicaciones sobre el conjunto de datos DAWN, incluidos los métodos de conjunto [32], las modificaciones de YOLO [33], las arquitecturas basadas en GAN [34], el transformador LDETR [35] y YOLO mejorado con algoritmos metaheurísticos [36]. . • En particular, DAWN es un conjunto de datos muy desafiante, como lo corrobora nuestra propia experiencia y lo subrayan las observaciones de los autores en [33], quienes comentaron: "Encontramos que el conjunto de datos DAWN es un poco más desafiante que los demás". Este desafío se debe al hecho de que algunas imágenes y objetos se caracterizan por una visibilidad extremadamente baja, lo cual es un factor que puede afectar la puntuación resultante de cualquier modelo desarrollado. • El dominio de la detección de objetos en condiciones climáticas adversas aún requiere conjuntos de datos más confiables que proporcionen suficiente variabilidad para cubrir diversas apariencias de objetos, condiciones de iluminación y oclusiones. La creación de dichos conjuntos de datos requiere mucho tiempo y es costosa. Un artículo publicado recientemente por Liu et al. [37] demostraron un enfoque basado en simulador que permite una fácil manipulación de las condiciones ambientales, la ubicación de los objetos y las perspectivas de la cámara. El uso de la recopilación de datos basada en simuladores abre la puerta a conjuntos de datos diversos y completos sin una recopilación extensa de datos del mundo real. Este enfoque puede acelerar la recopilación de datos al configurar y ejecutar varios escenarios climáticos adversos sin, por ejemplo, esperar los cambios estacionales del clima. Además , ofrece escalabilidad de datos, superando las limitaciones geográficas de la recopilación de datos del mundo real.

• Si bien las publicaciones recientes y los conjuntos de datos públicos existentes ofrecen recursos valiosos para la detección de objetos en diversas condiciones climáticas, existe una clara necesidad de realizar más trabajo que incluya escenarios de clima arenoso. • Combinar imágenes con LiDAR mediante fusión puede ser un enfoque prometedor para mejorar la detección de objetos en entornos de vehículos autónomos. Estudios recientes, como los de Dai et al. [38] y Liu et al. [39], han demostrado que esta técnica mejora significativamente la detección de objetos en entornos desafiantes al aprovechar las características complementarias tanto de LiDAR como de las cámaras. Las cámaras proporcionan una solución liviana y rentable que captura ricos detalles de colores y texturas, ayudando en la clasificación e identificación de objetos. Por otro lado, LiDAR ofrece mediciones de distancia precisas e información espacial 3D, que son particularmente útiles en condiciones de baja visibilidad donde las cámaras pueden tener dificultades. Al fusionar los datos de ambos sensores, se puede mejorar enormemente la precisión y solidez de los sistemas de detección de objetos. • Ampliamos nuestros experimentos para probar nuestro modelo utilizando el conjunto de datos UAVDT [40]. El conjunto de datos original del UAVDT comprende más de 77.000 imágenes capturadas durante el día, la noche y en condiciones climáticas de niebla. Después de ejecutar el experimento durante 64 épocas, logramos los siguientes resultados: mAP del 94,1%, recuperación del 90,8% y precisión del 97,0%. Creemos que el conjunto de datos UAVDT requiere un preprocesamiento adicional antes de que pueda utilizarse por completo. Por ejemplo, ajustar el marco temporal para capturar imágenes podría ayudar a diversificar las imágenes resultantes.

• Los datos sintéticos se pueden utilizar para abordar los desafíos y limitaciones de los conjuntos de datos del mundo real. En una publicación reciente [41], los autores propusieron CrowdSim2, un conjunto de datos sintéticos, para tareas de detección de objetos, en particular la detección de personas y vehículos. Esta técnica puede ser muy beneficiosa para el dominio AV al ofrecer un entorno controlado donde se pueden controlar factores como las condiciones climáticas, la densidad de objetos y la iluminación.

manipulado con precisión, lo que permite probar modelos de detección de objetos en varios escenarios. Además, se puede utilizar para simular eventos raros pero críticos, como accidentes u obstáculos inusuales, que pueden estar subrepresentados en conjuntos de datos del mundo real.

9. Conclusiones

Clasificar el clima y detectar objetos en entornos climáticos severos es una tarea crítica y desafiante. En este artículo presentamos un modelo de objetivos múltiples que integra la clasificación del clima y la detección de objetos y los trata como un problema unificado dentro del dominio de los sistemas de percepción de vehículos autónomos. Nuestro modelo consta de dos bloques principales. Primero, el bloque de calidad verifica la calidad de la imagen según la puntuación BRISQUE y, si la imagen tiene una puntuación superior al umbral, continúa mejorando mediante un método SRGAN. En segundo lugar, el bloque Clasificar y Detectar clasifica cuatro tipos de clima adverso (nieve, lluvioso, niebla y arena) y detecta seis clases (automóvil, ciclista, peatón, motocicleta, autobús y camión). Durante nuestro desarrollo, utilizamos el desafiante conjunto de datos DAWN como nuestra fuente de imágenes y empleamos YOLO como estructura base para la clasificación y detección. Los resultados experimentales muestran que nuestro modelo logró una precisión promedio promedio (mAP) del 74,6% para detectar objetos utilizando la arquitectura YOLO con la arquitectura C3 como columna vertebral y SiLU como función de activación.

Además, para clasificar el clima de la escena, nuestro modelo alcanzó una precisión del 74,3%, lo que se alinea estrechamente con el mapa. Dicho esto, todavía existen algunos desafíos en el dominio que deben considerarse al desarrollar modelos de detección y clasificación. Los cambios en las características de la escena, como la iluminación y la nubosidad, conducen a una clasificación errónea del tiempo correcto.

10. Trabajo futuro

Las condiciones climáticas adversas siguen siendo un ámbito muy desafiante en los entornos audiovisuales. Para lograr el más alto nivel de automatización, los sensores de las cámaras necesitan un sistema robusto que sea capaz de navegar de forma segura en diversos escenarios meteorológicos y capaz de observar con precisión el entorno. En el futuro, ampliaremos nuestro dominio para incluir conjuntos de datos adicionales que podrían fusionarse con el conjunto de datos actual de DAWN. Esto podría llevarnos a ampliar nuestras clases de detección para incluir clases nuevas y más detalladas que observemos en un entorno de conducción real, como semáforos, niños, animales domésticos (como perros) y personal encargado de hacer cumplir la ley (como agentes de policía). Cada una de estas clases representa componentes integrales de la escena de la carretera, y detectar y responder con precisión a su presencia es esencial para garantizar la seguridad y eficiencia de los sistemas de conducción autónoma. Al incorporar estas clases adicionales en nuestro marco de detección, nuestro objetivo es mejorar el mAP general. Además, nuestro objetivo es mejorar las capacidades de percepción de los sistemas autónomos mediante la fusión perceptual, que implica combinar información de múltiples sensores, como cámaras, LiDAR, radar y sensores ultrasónicos, para crear una representación completa y precisa del entorno circundante. Al desarrollar un sistema tan robusto , creemos que podemos mitigar el impacto de las condiciones climáticas adversas en el rendimiento de los sensores y mejorar la confiabilidad y solidez de los sistemas generales de percepción AV.

Contribuciones del autor: Software NA; análisis de datos, NA, AA y AB; curación de datos, NA; metodología, NA; redacción: preparación del borrador original, NA; redacción: revisión y edición, AA y AB; supervisión, AA y AB; conceptualización, NA, AA y AB Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: Esta investigación no recibió financiamiento externo.

Declaración de disponibilidad de datos: los datos están contenidos en el artículo.

Conflictos de intereses: Los autores declaran no tener conflictos de intereses.

Referencias

1. J3016_202104; Taxonomía y definiciones de términos relacionados con los sistemas de automatización de la conducción de vehículos de motor de carretera. Sociedad de Ingenieros Automotrices (SAE): Warrendale, PA, EE. UU., 2021.
2. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich presenta jerarquías para una detección precisa de objetos y una segmentación semántica. En Actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones, Columbus, OH, EE. UU., 23 a 28 de junio de 2014; págs. 580–587.
3. Girshick, R. R-CNN rápido. En Actas de la Conferencia Internacional IEEE sobre Visión por Computador, Sevilla, España, 17-19 de marzo de 2015; págs. 1440-1448.
4. Ren, S.; Él, K.; Girshick, R.; Sun, J. Faster R-CNN: Hacia la detección de objetos en tiempo real con redes de propuesta de región. Adv. Inf. neuronal . Proceso. Sistema. 2015, 28, 91–99. [Referencia cruzada] [PubMed]
5. Él, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Máscara R-CNN. En Actas de la Conferencia Internacional IEEE sobre Visión por Computadora, Jaipur, India, 18 a 21 de diciembre de 2017; págs. 2961–2969.
6. Lin, TY; Dollár, P.; Girshick, R.; Él, K.; Hariharan, B.; Belongie, S. Funciones de redes piramidales para la detección de objetos. En Actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones, Penang, Malasia, 5 a 8 de noviembre de 2017; págs. 2117-2125.
7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. Solo miras una vez: detección de objetos unificada en tiempo real. En Actas de las actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones, Bangalore, India, 20 y 21 de mayo de 2016; págs. 779–788.
8. Redmon, J.; Farhadi, A. Yolov3: Una mejora incremental. arXiv 2018, arXiv:1804.02767.
9. Wang, CY; Bochkovskiy, A.; Liao, HYM YOLOv7: La bolsa de regalos entrenable establece un nuevo estado del arte para objetos en tiempo real detectores. arXiv 2022, arXiv:2207.02696.
10. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Caña, S.; Fu, CY; Berg, AC Ssd: detector multibox de disparo único. En Actas de la Conferencia Europea sobre Visión por Computador, Ámsterdam, Países Bajos, 11 a 14 de octubre de 2016; Springer: Berlín/Heidelberg, Alemania, 2016; págs. 21–37.
11. Dhananjaya, MM; Kumar, VR; Yogamani, S. Clasificación climática y de nivel de luz para la conducción autónoma: conjunto de datos, línea de base y aprendizaje activo. En Actas de la Conferencia Internacional de Sistemas de Transporte Inteligentes (ITSC) IEEE de 2021, Indianápolis, IN, EE. UU., 19 a 22 de septiembre de 2021; IEEE: Piscataway, Nueva Jersey, EE. UU., 2021; págs. 2816–2821.
12. Kondapally, M.; Kumar, KN; Visnú, C.; Mohan, CK Hacia un enfoque de reconocimiento de escenas meteorológicas de transición para Vehículos Autónomos. Traducción IEEE. Intel. Transp. Sistema. 2023, 25, 5201–5210.
13. Ibrahim, señor; Haworth, J.; Cheng, T. WeatherNet: Reconocimiento de las condiciones meteorológicas y visuales a partir de imágenes a nivel de calle utilizando aprendizaje residual profundo. ISPRS Int. J. Geo-Inf. 2019, 8, 549. [Referencia cruzada]
14. Xia, J.; Xuan, D.; Bronceado, L.; Xing, L. ResNet15: Reconocimiento del tiempo en carreteras con red neuronal convolucional profunda. Adv. Meteorol. 2020, 2020, 6972826. [Referencia cruzada]
15. Él, K.; Zhang, X.; Ren, S.; Sun, J. Aprendizaje residual profundo para el reconocimiento de imágenes. En las actas de la Conferencia IEEE sobre Visión por computadora y reconocimiento de patrones, Bangalore, India, 20 y 21 de mayo de 2016; págs. 770–778.
16. Cao, Y.; Li, C.; Peng, Y.; Ru, H. MCS-YOLO: un método de detección de objetos multiescala para un entorno vial de conducción autónoma reconocimiento. Acceso IEEE 2023, 11, 22342–22354.
17. Liu, Z.; Lin, Y.; Cao, Y.; Eh.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Transformador Swin: transformador de visión jerárquica que utiliza ventanas desplazadas. En actas de la Conferencia internacional IEEE/CVF sobre visión por computadora, Montreal, BC, Canadá, 11 a 17 de octubre de 2021; págs. 10012-10022.
18. Pan, G.; Fu, L.; Yu, R.; Muresan, MI Reconocimiento del estado de la superficie de la carretera en invierno mediante una red neuronal convolucional profunda previamente entrenada . En actas de la 97.ª reunión anual de la Junta de Investigación del Transporte, Washington, DC, EE. UU., 7 a 11 de enero de 2018; págs. 838–855.
19. Wang, R.; Zhao, H.; Xu, Z.; Ding, Y.; Li, G.; Zhang, Y.; Li, H. Detección de objetivos de vehículos en tiempo real en condiciones climáticas adversas basado en YOLOv4. Frente. Neurorobótica 2023, 17, 34.
20. Li, X.; Wu, J. Extracción de datos de movimiento de vehículos de alta precisión a partir de vídeos de vehículos aéreos no tripulados capturados en diversas condiciones climáticas. Sensores remotos 2022, 14, 5513. [CrossRef]
21. Liu, W.; Ren, G.; Yu, R.; Guo, S.; Zhu, J.; Zhang, L. YOLO adaptable a imágenes para la detección de objetos en condiciones climáticas adversas. En Actas de la Conferencia AAAI sobre Inteligencia Artificial, Filadelfia, PA, EE. UU., 27 de febrero a 2 de marzo de 2022; Volumen 36, págs. 1792–1800.
22. Huang, Carolina del Sur; Le, TH; Jaw, DW DSNet: Aprendizaje semántico conjunto para la detección de objetos en condiciones climáticas adversas. Traducción IEEE. Patrón Anal. Mach. Intel. 2020, 43, 2623–2633.
23. Sharma, T.; Debaque, B.; Duclos, N.; Chehri, A.; Más amable, B.; Fortier, P. Detección de objetos y percepción de escenas basada en aprendizaje profundo bajo malas condiciones climáticas. Electrónica 2022, 11, 563. [CrossRef]
24. Jung, HK; Choi, GS Yolov5 mejorado: detección eficiente de objetos utilizando imágenes de drones en diversas condiciones. Aplica. Ciencia. 2022, 12, 7255. [Referencia cruzada]
25. ABDULGHANI, AMA; DALVEREN, GGM Detección de objetos en movimiento en video con algoritmos YOLO y R-CNN más rápido en diferentes condiciones. Avrupa Bilim Ve Teknoloji Dergisi 2022, 33, 40–54. [Referencia cruzada]
26. Kenk, MA; Hassaballah, M. DAWN: Detección de vehículos en un conjunto de datos de naturaleza climática adversa. arXiv 2020, arXiv:2008.05402.
27. Mittal, A.; Moorthy, Alaska; Bovik, AC Evaluación de la calidad de la imagen sin referencia en el dominio espacial. Traducción IEEE. Proceso de imagen. 2012, 21, 4695–4708. [Referencia cruzada] [PubMed]

28. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al.
Superresolución fotorrealista de una sola imagen utilizando una red generativa adversaria. En Actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones, Penang, Malasia, 5 a 8 de noviembre de 2017; págs. 4681–4690.
29. Zoph, B.; Cubuk, ED; Ghiasi, G.; Lin, TY; Shlens, J.; Le, QV Aprendizaje de estrategias de aumento de datos para la detección de objetos.
En Actas de Computer Vision – ECCV 2020: 16.ª Conferencia Europea, Glasgow (Reino Unido), 23 a 28 de agosto de 2020; Springer: Berlín/Heidelberg, Alemania, 2020; págs. 566–583, Parte XXVII 16.
30. Volk, G.; Müller, S.; Von Bernuth, A.; Hospach, D.; Bringmann, O. Hacia una detección robusta de objetos basada en CNN mediante aumento con variaciones de lluvia sintéticas. En actas de la Conferencia sobre sistemas de transporte inteligentes (ITSC) de IEEE de 2019; IEEE: Piscataway, Nueva Jersey, EE. UU., 2019; págs. 285–292.

31. Parque, H.; Yoo, Y.; Seo, G.; Mano.; Yun, S.; Kwak, N. C3: Convolución concentrada-integral y su aplicación a la semántica
segmentación. arXiv 2018, arXiv:1812.04920.
32. Walambe, R.; Marathe, A.; Kotecha, K.; Ghinea, G. Marco de conjunto de detección de objetos livianos para vehículos autónomos en
condiciones climáticas desafiantes. Computadora. Intel. Neurociencias. 2021, 2021, 5278820. [Referencia cruzada] [PubMed]
33. Farid, A.; Hussain, F.; Khan, K.; Shahzad, M.; Khan, U.; Mahmood, Z. Un método de detección de vehículos en tiempo real rápido y preciso
utilizando el aprendizaje profundo para entornos sin restricciones. Aplica. Ciencia. 2023, 13, 3059. [Referencia cruzada]
34. Musat, V.; Fursa, I.; Newman, P.; Cuzzolin, F.; Bradley, A. Ciudad multiclima: acumulación de condiciones climáticas adversas para la conducción autónoma.
En Actas de las actas de la Conferencia internacional IEEE/CVF sobre visión por computadora, Montreal, BC, Canadá, 11 a 17 de octubre de 2021; págs. 2906–2915.

35. Tiwari, Alaska; Pattanaik, M.; Sharma, G. Transformador de detección de poca luz (LDETR): detección de objetos en condiciones de poca luz y condiciones adversas.
las condiciones climáticas. Multimed. Herramientas Aplica. 2024, 1–18. [Referencia cruzada]
36. Özcan, I.; Altún, Y.; Parlak, C. Mejora del rendimiento de detección YOLO de vehículos autónomos en condiciones climáticas adversas
Utilizando algoritmos metaheurísticos. Aplica. Ciencia. 2024, 14, 5841. [Referencia cruzada]
37. Liu, S.; Zhang, H.; Qi, Y.; Wang, P.; Zhang, Y.; Wu, Q. Aerialvln: Navegación por visión y lenguaje para vehículos aéreos no tripulados. En Actas de las actas de la Conferencia internacional IEEE/CVF sobre visión por computadora, París, Francia, 2 a 6 de octubre de 2023; págs. 15384-15394.
38. Dai, Z.; Guan, Z.; Chen, Q.; Xu, Y.; Sun, F. Detección mejorada de objetos en vehículos autónomos mediante LiDAR: sensor de cámara
Fusión. Electricidad mundial. Veh. J. 2024, 15, 297. [Referencia cruzada]
39. Liu, H.; Wu, C.; Wang, H. Detección de objetos en tiempo real mediante LiDAR y fusión de cámaras para conducción autónoma. Ciencia. Rep. 2023, 13, 8056. [CrossRef] [PubMed]

40. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. El punto de referencia de los vehículos aéreos no tripulados: detección y seguimiento de objetos. En Actas de la conferencia europea sobre visión por computadora (ECCV), Múnich, Alemania, 8 a 14 de septiembre de 2018; págs. 370–386.

41. Foszner, P.; Szczesna, A.; Ciampi, L.; Mesina, N.; Cygan, A.; Bizón', B.; Cogiel, M.; Golba, D.; Macioszek, E.; Staniszewski, M.
CrowdSim2: un punto de referencia sintético abierto para detectores de objetos. arXiv 2023, arXiv:2304.05090.

electrónica

MDPI

Artículo

Impacto de la sobrecarga del enlace de comunicación en el flujo de energía y Transmisión de datos en sistemas de energía ciberfísicos

Xinyu Liu 1,2,3, Yan Li 1,2,3,* y Tianqi Xu 1,2,3

1 Escuela de Tecnología Eléctrica y de la Información, Universidad Yunnan Minzu, Kunming 650504, China; xyliu7918@gmail.com (XL); xu.tianqi@ymu.edu.cn (Texas)
2 Laboratorio clave de sistema autónomo no tripulado de Yunnan, Kunming 650504, China
3 Laboratorio clave de sistema de energía ciberfísica de colegios y universidades de Yunnan, Kunming 650504, China
* Correspondencia: yan.li@ymu.edu.cn

Resumen: El volumen de demanda de flujo en los sistemas de energía ciberfísicos (CPPS) fluctúa de manera desigual entre las redes acopladas y es susceptible a congestión o sobrecarga debido a la demanda de energía de los consumidores o desastres extremos. Por tanto, considerando la elasticidad de las redes reales, los enlaces de comunicación con un flujo excesivo de información no se desconectan inmediatamente sino que tienen un cierto grado de redundancia. Este artículo propone un modelo iterativo dinámico de fallas en cascada basado en la distribución de la sobrecarga del flujo de información en una red de comunicación y la intermediación del flujo de energía en la red eléctrica física. Primero, se introduce un modelo de capacidad de carga no lineal de una red de comunicación con sobrecarga y bordes ponderados, considerando completamente los tres estados del enlace: normal, falla y sobrecarga. Luego, la intermediación de flujo sustituye a los flujos ramales en la red eléctrica física, y el flujo de energía en las líneas fallidas se redistribuye utilizando el modelo de capacidad de carga, simplificando los cálculos. En tercer lugar, bajo la influencia de las relaciones de acoplamiento, se construye un modelo integral basado en una teoría de percolación mejorada, con estrategias de ataque formuladas para evaluar con mayor precisión las redes acopladas. Las simulaciones en el sistema de bus IEEE-39 demuestran que considerar la capacidad de sobrecarga de los enlaces de comunicación a pequeña escala mejora la robustez de las redes acopladas. Además, los ataques deliberados a enlaces causan daños más rápidos y extensos en comparación con los ataques aleatorios.

Cita: Liu, X.; Li, Y.; Xu, T. Impacto de sobrecarga del enlace de comunicación en Flujo de energía y transmisión de datos en Sistemas de energía ciberfísicos. Electrónica 2024, 13 , 3065. https://doi.org/10.3390/electronics13153065

Editor académico: Christos J. Bouras

Recibido: 8 de julio de 2024
Revisado: 30 de julio de 2024
Aceptado: 31 de julio de 2024
Publicado: 2 de agosto de 2024

Palabras clave: bordes sobrecargados; flujo de información; flujo de energía; fallas en cascada; teoría de percolación mejorada ; sistemas de energía ciberfísicos

1. Introducción

1.1. Antecedentes

Con el desarrollo de la red inteligente y la Internet energética, el sistema eléctrico se ha acoplado profundamente al sistema de información. El sistema de energía en el lado físico y el sistema de comunicación en el lado de la información han evolucionado gradualmente hasta convertirse en el sistema de energía ciberfísico (CPPS) [1]. Si bien el sistema acoplado ha aportado numerosos beneficios, también ha aumentado el riesgo de fallos en cascada en todo el espacio. Las vulnerabilidades en los dos sistemas a través de una red superpuesta aumentarán el riesgo de propagación de fallas, de modo que incluso una falla de un solo borde o nodo puede afectar a toda la red, lo que a menudo conduce a un colapso global [2,3]. Por ejemplo, el apagón masivo en el oeste de Estados Unidos en 2003, el apagón en Ucrania en 2015 y el apagón 815 en Brasil en 2023 [4,5] fueron todos causados por fallas en bordes específicos de la red de información. Estas fallas se propagaron a la red eléctrica a través del acoplamiento funcional, lo que finalmente resultó en la parálisis simultánea de ambos sistemas.

1.2. Obras relacionadas

El análisis de casos pasados muestra que cuando el sistema de comunicación falla o es atacado, los paquetes de datos pueden perderse o manipularse, impidiendo el control de circuito cerrado. Debido a la conexión de acoplamiento ciber-físico, la falla se propagará para afectar los nodos de energía en la red física con cierta probabilidad. Luego, la falla continúa propagándose en la red física y, en última instancia, causa graves daños al sistema [6–8]. Por lo tanto, modelar la red eléctrica física, la red de comunicaciones y su conexión de acoplamiento es esencial para comprender el proceso de propagación de fallas en cascada a través del espacio.

En [9], se propuso un modelo de acoplamiento uno a uno entre nodos de energía y comunicación , analizando la robustez del sistema de falla en cascada después de la eliminación de una pequeña fracción de nodos según su modelo topológico. En [10], los autores estudiaron la robustez de una red de comunicación sin escala de doble capa basada en la teoría de la percolación. Basado en [10], ref. [11] consideraron las interacciones entre nodos en diferentes capas como heterogéneas, estudiando un tipo de dinámica en cascada en redes de doble capa que exhiben tanto interdependencia como conectividad. Basado en [9,10], Chen et al. [12] diferenciaron los nodos de la red eléctrica física en nodos generadores y de carga, proponiendo un nuevo mecanismo interactivo para fallas en cascada. En [13], las características operativas y la estructura topológica de la red de transmisión se integraron para establecer un modelo de falla en cascada para fallas aleatorias en líneas de transmisión bajo diferentes estrategias de acoplamiento, con el objetivo de obtener una red acoplada robusta óptima. Sin embargo, los modelos de acoplamiento establecidos en estos estudios se centran únicamente en estructuras topológicas, descuidando las características operativas de ambos lados de las redes acopladas. En [14,15], se consideró la optimización del flujo de energía en la red eléctrica física y se compararon los resultados de vulnerabilidad bajo diferentes estrategias y topologías de la red de información, pero no se consideraron las características operativas de la red de información. En [16], se estudió la propagación dinámica de fallas en cascada entre la red eléctrica y la red de comunicaciones, considerando las características del flujo de energía y el flujo de datos en dos sistemas diferentes, pero el impacto de la sobrecarga de datos en la red de comunicaciones en la red acoplada. no fue considerado.

En [17], se estudiaron las características de recuperación de diferentes fuerzas de acoplamiento y topologías de red basadas en un modelo en cascada relacionado con la carga. Aunque en este modelo se consideró el estado de sobrecarga de los nodos , la carga adicional no se redistribuyó. Basado en [16,17], Ding et al. [18] propusieron un modelo mejorado de fallas en cascada. Este modelo considera el estado de sobrecarga y el proceso de recuperación de los cibernodos, así como la optimización del flujo de energía en la capa física y la redistribución del flujo de información durante la propagación de fallas. Sobre esta base, Wang et al. [19] emplearon un modelo de flujo de energía de CA para caracterizar las características operativas de la red eléctrica, mejorando la precisión del modelo de red eléctrica. Simultáneamente, construyó una red de comunicación ponderada con centros de control y aplicó un modelo de redistribución de flujo. En [20], los autores propusieron dos tipos de modelos de dependencia fuerte y débil y analizaron los cambios de robustez de la red de acoplamiento utilizando un esquema de equilibrio de carga consciente de la congestión bajo fallas aleatorias iniciales en la capa de energía. Sin embargo, no se consideraron los flujos de datos en la capa de comunicación. Los autores de [21,22] consideraron fallas en los nodos de comunicación y establecieron un modelo mejorado de fallas en cascada basado en la distribución de carga de la capa física. En [23], el modelo propuesto por los autores consideró las diferencias prácticas entre una red de comunicación y una red eléctrica en términos de estructura de red, operación física y comportamiento dinámico, enfocándose en analizar fallas que ocurren en el lado de la red eléctrica. De estos estudios se desprende que la mayoría de los estudiosos han prestado menos atención a los retrasos en la transmisión causados por sobrecargas de tráfico en las redes de comunicación y al establecimiento de modelos acoplados que incorporan las características operativas de ambas redes.

1.3. Motivación

De hecho, muchos bordes conectados suelen poseer capacidad redundante. Por ejemplo, la Figura 1a ilustra una red de comunicación con cinco nodos. La matriz F representa la matriz de demanda de transmisión de flujo de información, donde los elementos Fij indican el flujo de información.

demanda que debe transmitirse desde el origen al destino. Cada enlace eij está asociado con su atributo de calidad qij [24,25], donde el nivel operativo de un enlace eij se define como la relación entre la capacidad del enlace y la carga del enlace. Suponiendo que el umbral de falla del enlace es ρ, si las fallas iniciales en la red de comunicación son enlaces con ρ < 0,5, estos enlaces se eliminarán de la red original (por ejemplo, elimine 1→2, 2→3). En este punto, la transmisión del flujo de información en el enlace de comunicación no se ve afectada (los enlaces 1→4→2 y 2→1→3 todavía existen), pero los enlaces 1→4 y 2→1 están en un estado de sobrecarga. La demanda de flujo de información en el enlace original 1→2 se redistribuirá al enlace 1→4→2. Aunque la red no se ve afectada inmediatamente y los enlaces sobrecargados no fallan, la calidad de la transmisión seguirá disminuyendo. Al aumentar el umbral a ρ = 0,7, la eficiencia operativa del enlace 1→4 cae por debajo del umbral crítico, lo que provoca que el estado del borde cambie de sobrecargado a fallido y se elimine de la red original. En este punto, se ven afectadas 25 demandas de transmisión de flujo de información (resaltadas en rojo en la figura).

Cuando el umbral se incrementa aún más hasta ρ = 0,9, se eliminan los vínculos con qij < 0,9. Como se muestra en la Figura 1d, sólo once unidades de demanda de tráfico pueden transmitirse efectivamente al centro de control. Se puede observar que durante el proceso de cambio de umbral y eliminación de enlaces, si no se consideran el estado de sobrecarga y la demanda de flujo de transmisión de los enlaces , la red colapsará prematuramente, provocando pérdidas importantes en la red eléctrica.

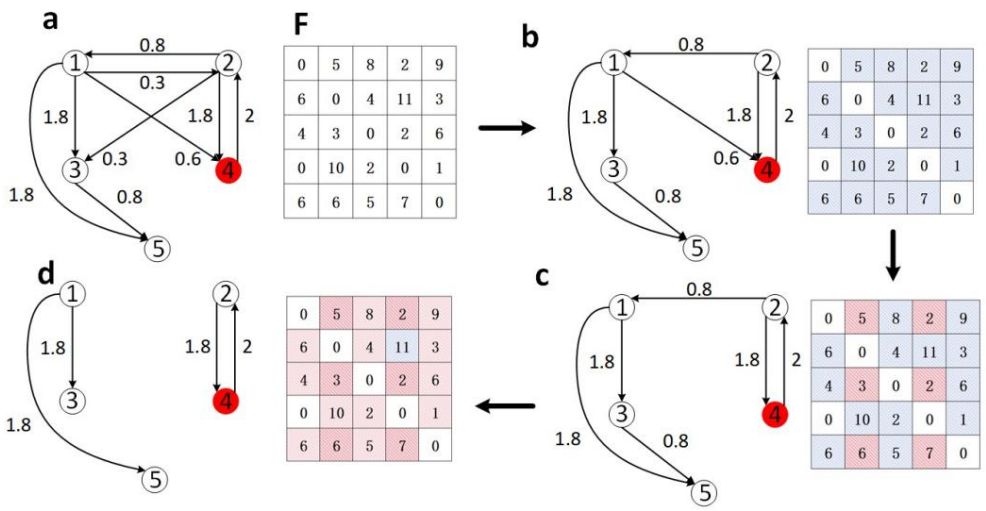


Figura 1. Diagrama de falla de enlace de la red de comunicaciones eléctricas en función de la demanda de flujo de información. (a) Red de comunicación G con tamaño n = 5, donde el nodo cuatro es el centro de control y los demás son nodos de transmisión regulares. La calidad qij de cada línea eij está marcada al lado del enlace. La matriz F representa la demanda de transmisión del flujo de información. (b) Suponiendo que las fallas iniciales en la red de comunicación son enlaces con ρ < 0,5. (c) Aumentar el umbral a ρ = 0,7, eliminando al mismo tiempo los vínculos con qij < 0,7. (d) Cuando el umbral se aumenta aún más a ρ = 0,9, los enlaces con qij < 0,9 se eliminan y sólo once unidades de demanda de tráfico pueden transmitirse efectivamente al centro de control.

Normalmente, cuando ocurre una falla N-1 en la red eléctrica, se debe volver a calcular la convergencia del flujo para cada escenario. Si el flujo converge, se debe determinar si el flujo en cada línea excede sus límites; si es así, se debe cortar la línea afectada. Si no converge, normalmente se implementan ajustes de salida del generador o deslastre de carga para equilibrar los flujos de energía. Enumerar todos los escenarios puede resultar complejo y llevar mucho tiempo.

Por lo tanto, este artículo propone aplicar la intermediación de flujo de línea [26] al modelo de capacidad de carga del sistema eléctrico, utilizándolo como carga de potencia en la línea. Este enfoque refleja y cuantifica efectivamente el papel de las líneas en la transmisión de energía desde los generadores a las cargas, considerando también el impacto de la máxima potencia de transmisión disponible entre generadores y cargas en las líneas críticas. Este entorno físico se alinea más estrechamente

el funcionamiento real de los sistemas eléctricos y refleja mejor sus características operativas. Al modelar el proceso de transmisión del flujo de información, distinguirlo de la red eléctrica es crucial. Los enlaces de la capa de información no se desconectarán debido a una sobrecarga del flujo de información, pero causarán congestión si el flujo es excesivo. Es necesario tener en cuenta el estado de sobrecarga al redistribuir el flujo de información desde enlaces no operativos a enlaces vecinos para desarrollar un modelo de flujo de red de energía-comunicaciones más realista . Por lo tanto, este artículo, considerando las interrupciones de enlaces en el sistema de comunicación, propone un modelo de falla en cascada de capacidad de carga basado en la teoría de percolación mejorada, integrando la sobrecarga del flujo de información y las características operativas del sistema de energía. Este modelo analiza el impacto de los enlaces sobrecargados en la robustez del sistema.

1.4. Contribuciones

Las contribuciones de este artículo se resumen a continuación:

- Teniendo en cuenta el estado de congestión de los enlaces de la red de comunicaciones, se establece un modelo de asignación de transmisión dinámica para el flujo de información en tres estados de enlace: normal, sobrecarga y falla. Se proponen métricas para la integridad de la topología del sistema de comunicación y las características operativas para evaluar la vulnerabilidad del sistema en caso de fallas.

- Considerando las características de generación de topología de las redes de comunicación de energía actuales , se propone una teoría de percolación mejorada. Los nodos de comunicación que inicialmente están fuera del componente conectado más grande pero que tienen enlaces de comunicación con el centro de control se consideran nodos efectivos. Los nodos de energía que pierden acoplamiento con la red de comunicación pero siguen siendo autoconsistentes también se consideran nodos efectivos.

- Considerando las características físicas de las redes acopladas, el indicador de interconexión de flujo de línea se utiliza para medir las características eléctricas de los flujos de energía de línea. Se propone un modelo de distribución de capacidad de carga basado en la intermediación de flujo para transferir la carga de líneas fallidas.

El resto del documento está organizado de la siguiente manera. En la Sección 2, construimos el modelo de dependencia unidireccional de CPPS y proporcionamos una descripción detallada del proceso de propagación de fallas a través del espacio considerando las condiciones de sobrecarga del enlace de comunicación dentro del modelo acoplado. En la Sección 3, proponemos dos métricas de evaluación de la robustez del sistema para medir los resultados de fallas en cascada. En la Sección 4, se realiza la simulación numérica para analizar la falla en cascada y presenta discusiones relacionadas. La sección 5 concluye el artículo.

2. Métodos

2.1. Modelado de redes acopladas de comunicación y energía basado en dependencia unidireccional

CPPS incluye el modelado de la red de información, la red de energía y la capa de acoplamiento. La red de información comprende la capa de acceso, la capa troncal y la capa cen
Los bordes interdependientes entre los nodos de energía y los nodos de información forman la red de acoplamiento [27]. En la red de acoplamiento, la red eléctrica proporciona soporte de energía para la red de comunicación, mientras que la red de comunicación ofrece soporte 3C para la red eléctrica. Sin embargo, dado que los nodos de comunicación están ampliamente equipados con fuentes de energía de respaldo, la falla de los nodos de energía acoplados no causa la falla de los nodos de comunicación debido a un corte de energía. El funcionamiento normal de los nodos de energía depende del monitoreo y control de los nodos de comunicación. La falla de los enlaces de comunicación dará como resultado que el centro de control no pueda recibir la información de fallas del sistema de energía de manera oportuna, lo que provocará que no se puedan abordar con prontitud las fallas de energía y se ampliará el alcance del impacto de las fallas. Por lo tanto, este artículo estudia principalmente el modelo de dependencia de la red eléctrica de la red de información. Con referencia al diagrama de flujo de la Figura 2, el proceso de modelado se detalla a continuación:

(1) La red eléctrica física se resume como un gráfico Gp(Vp, Ep) compuesto por nodos de potencia Vp y líneas de transmisión Ep. El conjunto Vp incluye equipos físicos en la red eléctrica, como plantas de energía, subestaciones y cargas.

(2) Dividir la red eléctrica en regiones [28].

a. La partición de comunidades se aplica a la división de regiones de la red eléctrica. En la red eléctrica, cuanto más cercana es la distancia entre dos autobuses, menor es la reactancia de la línea y mayor es su recíproco (es decir, mayor peso), lo que indica una mayor intimidad entre el par de nodos. Por lo tanto, es más probable que dichos nodos se divida en la misma región. El recíproco de la reactancia de línea se utiliza como peso para los elementos distintos de cero de la matriz de adyacencia de la red eléctrica E. b. Utilizando el método Fast Newman, la matriz E modificada de los pasos anteriores se sustituye por la matriz original E para calcular la modularidad Q para la partición inicial de la red eléctrica. El proceso de partición debe satisfacer las siguientes condiciones: cada región debe contener al menos un generador y una carga; el número de regiones debe ser menor que el mínimo del número de generadores y cargas; y las subregiones deben lograr el equilibrio de poder. C. Para lograr el equilibrio de poder dentro de las regiones, los flujos de energía del sistema en las líneas se utilizan como valores de peso basados en el algoritmo Prim. Estos valores de peso se utilizan para determinar las rutas de flujo desde los generadores hasta las cargas. Luego, las regiones se fusionan según estas rutas y se combinan con los resultados de la partición inicial para formar las zonas finales.

(3) Se determina el número de nodos para cada capa: la capa de acceso, la capa principal y la capa central. Los nodos de la capa de acceso son nodos de recopilación de información, siendo el número de nodos de comunicación igual al número de nodos de energía; los nodos de la capa principal son nodos de equipos de enrutamiento, también considerados nodos de control, siendo el número de nodos igual al número de particiones de la red eléctrica; y los nodos de la capa central son dos nodos de control que representan el despacho principal y el de respaldo.

(4) El modelado de la red de información: a. La conexión entre la red eléctrica y la capa de acceso: de acuerdo con el diagrama abstracto de la red eléctrica Gp, los nodos de la capa de acceso están conectados de manera correspondiente uno a uno para que el diagrama de topología de la capa de acceso sea consistente con el diagrama de topología. de la red eléctrica. b. La conexión entre la capa de acceso y la capa principal: los nodos de mayor grado en cada región de la capa de acceso y los nodos generadores están conectados a los nodos correspondientes en la capa principal. C. Los nodos de la capa principal están conectados internamente de acuerdo con la conexión relación de las particiones de la red eléctrica.

d. La conexión entre la capa de acceso y la capa central: calcule la proporción qi de la salida Gi del generador con respecto a la salida total del sistema y conéctela a las llamadas principal y de respaldo con una probabilidad de qi .

mi. Todos los nodos de enrutamiento están conectados a los nodos del centro de programación principal y de respaldo. F. Los centros de programación principal y de respaldo están conectados.

(5) Resuma la red de información para formar Gc.

(6) Conecte los nodos de la capa de acceso con los nodos de potencia para formar bordes dependientes unidireccionales.

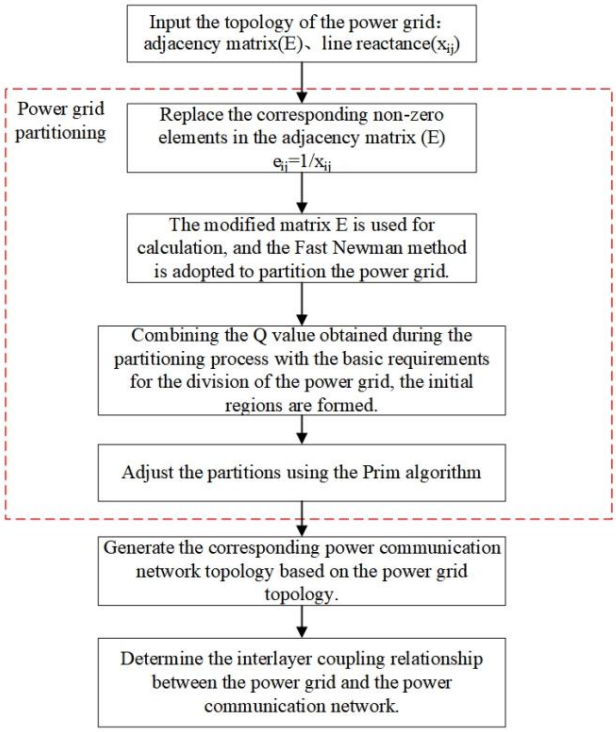


Figura 2. Diagrama de flujo del modelado de sistemas de energía ciberfísicos.

Así, las redes acopladas se muestran en la Figura 3.

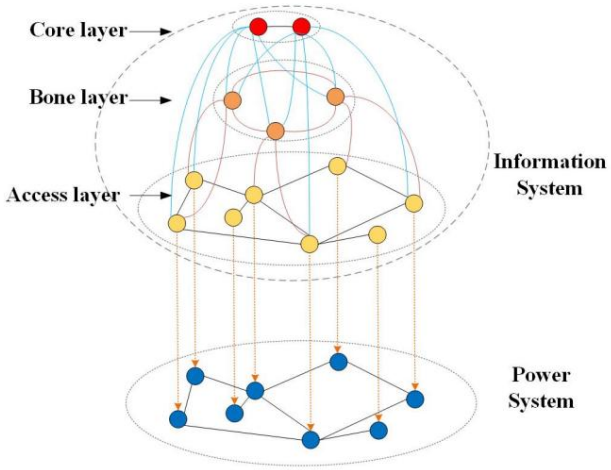


Figura 3. Diagrama de topología de la red eléctrica dependiente de la red de información.

2.2. Modelo de falla en cascada basado en la teoría de la percolación que considera sobrecargas en los enlaces de comunicación

2.2.1. Modelo de flujo no lineal para enlaces

El modelo de flujo de información de enlace de la red de comunicaciones emplea un modelo de capacidad-carga no lineal que tiene en cuenta el estado de sobrecarga [29]. En redes complejas, un grado de nodo más alto indica más conexiones con otros nodos. Por lo tanto, los valores de grado de los puntos finales se utilizan para calcular el flujo de información de cada enlace de capa de acceso. Cuanto mayor sea el valor de grado de un vínculo, mayor será el flujo de información a través de él. La relación de flujo de información de un enlace se utiliza como peso del enlace, y el coeficiente de sobrecarga δ describe la capacidad del borde para manejar flujo de información adicional, de la siguiente manera:

$$C_{c,ij} = L_{c,ij} + \beta L_{c,ij}^{\alpha} \quad (1)$$

$$C^{\text{máximo}}_{c,ij} = \delta Cc,ij \quad (2)$$

$$Wc,ij = \begin{cases} \frac{lc,ij}{CC,ij} & eij \quad CE \\ 0 & eij \quad / CE \end{cases} \quad (3)$$

dónde

$$Lc,ij = wij = kikj^{\theta} \quad (4)$$

donde Lc,ij representa la cantidad de información transmitida por el enlace y ki y kj denotan los grados de los nodos i y j, respectivamente. θ es un parámetro que ajusta el flujo de información. Cc,ij es la capacidad del enlace, $C^{\text{máximo}}_{c,ij}$ representa el caudal máximo que puede soportar la línea, y Wc,ij es el peso del borde eij. α y β son coeficientes de capacidad. Cuando α = 1, el modelo es lineal.

2.2.2. Teoría de percolación mejorada La

teoría de percolación tradicional [30] se usa ampliamente para describir la estructura, función y resiliencia de los sistemas de red. Los modelos de percolación simulan escenarios de falla de enlaces eliminando gradualmente enlaces de la red. A medida que los enlaces se eliminan progresivamente, la reducción del tamaño del subgrafo conectado más grande se puede utilizar para medir las consecuencias de las fallas de los enlaces. Por lo tanto, la teoría de la percolación es aplicable al modelado de fallas en cascada en sistemas de energía ciberfísicos. Sin embargo, la teoría de la percolación tradicional solo considera el subconjunto conectado más grande en una red de una sola capa al determinar el subconjunto de nodos de trabajo, sin considerar otros escenarios. Su aplicación directa para modelar fallas en cascada en sistemas de energía ciberfísicos dificulta la simulación precisa del proceso de falla. Teniendo en cuenta las características de transmisión del flujo de información de la red de información, donde la capa central y la capa principal no se corresponden directamente con la capa de acceso, y el estado del enlace considera la situación de sobrecarga, es necesario mejorar el modelo clásico de la teoría de la percolación. El modelo de falla del sistema de energía ciberfísico establecido en este artículo es el siguiente: • Sistema de comunicación. Para los enlaces en la capa de acceso, los bordes con pesos mayores que el coeficiente de sobrecarga se consideran

fallidos. Para los nodos de información en la capa de acceso, los nodos que no pueden establecer una ruta para controlar los nodos se consideran fallidos. • Sistema de poder. Un nodo de carga de energía debe estar

conectado a al menos un nodo generador; en caso contrario, se considera fallido. De manera similar, un nodo generador debe estar conectado a al menos un nodo de carga de energía; en caso contrario, se considera fallido. Los nodos que funcionan como generadores y cargas se consideran nodos autoconsistentes.

- Interacción. Los nodos de energía acoplados con nodos de comunicación fallidos también fallarán. Además, los nodos de energía acoplados con los nodos de comunicación que salen de los retrasos de transmisión fallarán con una cierta probabilidad Pdi.

2.2.3. Modelo de carga-capacidad de la red eléctrica basado en la intermediación de flujos

En estudios previos, el principio de propagación por el camino más corto se ha utilizado comúnmente para investigar la transmisión de energía entre buses en una red eléctrica. Sin embargo, en realidad, el flujo de energía en una red eléctrica no sigue sólo el camino con la impedancia más baja, sino que se propaga a lo largo de todos los caminos posibles, siguiendo la Ley de Kirchhoff.

Para reflejar con precisión el papel de cada línea de transmisión en la propagación de energía y el impacto variable de diferentes pares generador-carga en cada línea, considere que en una red eléctrica determinada, cada línea de transmisión transporta una proporción variable de potencia de transmisión P(m, n) desde el generador m hasta la carga n. En consecuencia, cada línea desempeña un papel distinto y tiene distintos niveles de importancia en la potencia de transmisión P (m, n). Dado que las rutas de transmisión de energía para los pares generador-carga que atraviesan cada línea difieren, la importancia de la línea dentro de todo el

La red se cuantifica utilizando el índice de intermediación de flujo [26]. Este índice se calcula considerando todos los pares generador-carga que utilizan la línea. La fórmula de cálculo es la siguiente:

$$FBij = \sum_{m \in G} \sum_{n \in L} \min(Sm, Sn) \frac{Pij,mPij,nPn}{PijPLnA \; unmPGm^{-1}} \quad (5)$$

donde G es el conjunto de nodos de generación y L es el conjunto de nodos consumidores. min(Sm, Sn) es el peso de la intermediación de flujo de una sola línea, que depende del valor mínimo entre la salida real del generador my la carga real n, lo que refleja la potencia de transmisión máxima disponible entre my n . Pij,m es la porción del flujo de potencia en la línea eij que se origina en el generador m. Pij,n es la porción del flujo de potencia en la línea eij dirigida hacia la carga n. Pn es el flujo de potencia del nodo de n. Pij es la potencia activa a través de la línea eij. PLn es la carga activa en el nodo de unm^{-1} contiene el orden inverso carga n. Una matriz de distribución de elementos. PGm es la salida activa del nodo generador m.

Por lo tanto, la carga de potencia Lp(ij) en el borde eij se puede calcular como

$$Lp(ij) = FBij \quad (6)$$

Considerando la capacidad de los enlaces para manejar cargas de energía, empleamos el modelo de capacidad de carga propuesto por Motter y Lai [31]. Según este modelo, la capacidad del enlace es directamente proporcional a su carga de potencia inicial, de la siguiente manera:

$$Cp(ij) = (1 + \gamma)Lp(ij) \quad (7)$$

donde γ es el parámetro de tolerancia.

2.2.4. Proceso de falla en cascada La red

de información tiene la capacidad de controlar nodos de energía. En consecuencia, las fallas que ocurren dentro de la red de información pueden propagarse a la red eléctrica interconectada. Esta sección detalla el proceso dinámico de fallas en cascada desencadenadas por ciertos enlaces de comunicación defectuosos.

Según la magnitud de la transmisión del flujo de información en la red de comunicaciones , los estados operativos de los enlaces se clasifican en tres tipos: normal, sobrecargado y fallido. Para analizar fallas en cascada entre redes acopladas, algunos enlaces en la red de comunicación se eliminan aleatoriamente como fallas iniciales. Con referencia al diagrama de flujo de la Figura 4, el proceso dinámico detallado de fallas en cascada provocadas por enlaces dañados específicos se describe a continuación:

- Paso 1: El conjunto de enlaces fallidos en la red de comunicación de la capa de acceso debido a fallas o ataques accidentales se denota como eij. El flujo de información en estos enlaces fallidos correrá a cargo de los bordes conectados a los nodos de los enlaces fallidos.
- Paso 2: El proceso de distribución del flujo de información sobre sucursales fallidas. Este artículo utiliza el principio de redistribución local del flujo de información, con la fórmula de cálculo proporcionada a continuación.

$$\Delta Lc,ia = Lc,ij \frac{Lc,ia}{\sum_{m \in \Omega 1} Lc,im + \sum_{n \in \Omega 2} Lc,jn} \quad (8)$$

donde Lc,ij es el flujo de información inicial en una rama eij, Lc,ia es el flujo de información inicial en una rama eia, Ω1 es el conjunto de nodos vecinos del nodo i, y Ω2 es el conjunto de nodos vecinos del nodo j.

• Paso 3: El proceso para determinar los estados sobrecargado y fallido es el siguiente.

$$\begin{cases} W_{c,im} > \delta & \text{fallar} \\ 1 < W_{c,im} < \delta \text{ y rand} > \text{sobrecarga de pim} & \\ 1 < W_{c,im} < \delta \text{ y rand} \le \text{pim fallan} & \\ W_{c,\text{soy}} \le 1 & \text{normal} \end{cases} \quad (9)$$

donde rand (0, 1).

Dado que cada rama tiene diferentes capacidades para manejar flujo de información adicional, se introduce un coeficiente de distribución ω para caracterizar esta propiedad.

$$\text{pía} = \left(\frac{W_{c,ia} - 1}{\delta - 1} \right)^{\omega} \quad (10)$$

• Paso 4: Proceso de distribución del flujo de información en sucursales sobrecargadas:

$$\Delta_{ak} = (L_{c,ia} - C_{c,ia}) T_{ak} \quad (11)$$

$$T_{ak} = \frac{C_{c,ak} - L_{c,ak}}{\sum_{e} \Omega_i (C_{c,ie} - L_{c,ie}) + \sum_{f} \Omega_a (C_{c,af} - L_{c,a\,f})} \quad (12)$$

donde Δ_{ak} es la estrategia de distribución para la rama sobrecargada, Ωi es el conjunto de nodos vecinos del nodo i con ramas en estado normal, y Ωa es el conjunto de nodos vecinos del nodo a con ramas en estado normal. • Paso 5: Determinar si hay nuevas ramas fallidas. Si hay nuevas ramas fallidas

detectado, continúe con el Paso 2; de lo contrario, continúe con el Paso 6.

• Paso 6: Cuente los conjuntos de enlaces fallidos y sobrecargados dentro de la red de comunicación . Actualice el conjunto efectivo de enlaces de comunicación y calcule los incrementos del retraso de transmisión para cada nodo de comunicación. Evaluar las condiciones de falla de los nodos de comunicación, eliminar los nodos que no pueden formar un camino hacia el centro de control y actualizar el conjunto de nodos de comunicación efectivos.

• Paso 7: Análisis de dependencia de control: Debido al acoplamiento uno a uno entre los nodos de comunicación de la capa de acceso y los nodos de energía, los nodos de comunicación fallidos identificados en el Paso 6 causarán que los nodos de energía correspondientes fallen a través de los bordes dependientes . Los nodos se seleccionan en función de la probabilidad de retraso de cada nodo de comunicación.

Si la información de estos nodos no se transmite al centro de control de manera oportuna, sus correspondientes nodos de energía se consideran fallidos. Marque los nodos de energía que han fallado en este paso. • Paso 8: Fallo físico: Primero, elimine los nodos

fallidos en la red eléctrica. A continuación, según las condiciones de falla de los nodos del sistema de energía, verifique si hay nodos fallidos adicionales entre los nodos restantes y elimínelos si es necesario. Recalcular la carga de energía en las líneas; Si la carga redistribuida excede la capacidad de cualquier línea, márquela como defectuosa. Repita este proceso hasta que no queden más nodos defectuosos en la red eléctrica. • Paso 9: Resultado: Finaliza la simulación de falla en cascada única. Genere los datos tanto para el

red eléctrica y la red de comunicaciones.

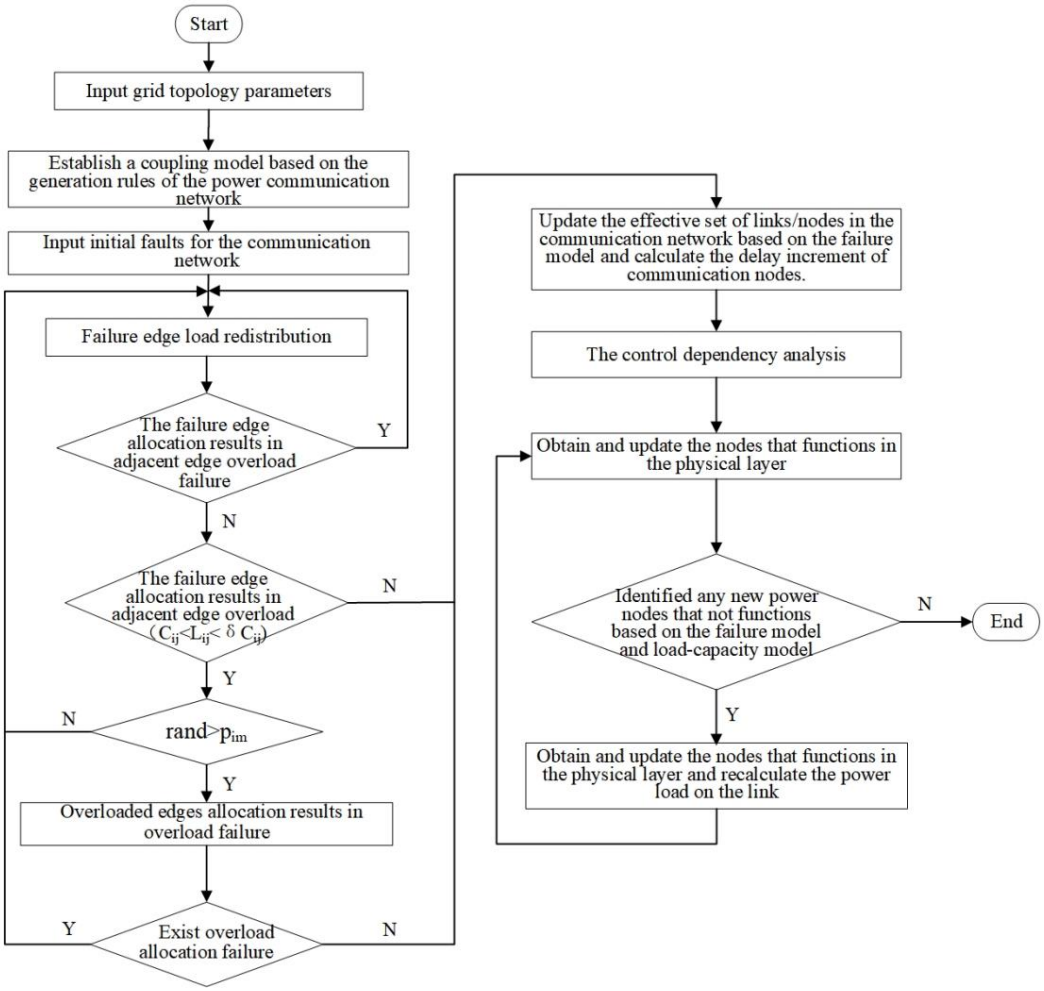


Figura 4. Diagrama de flujo del proceso de falla en cascada en sistemas de energía ciberfísicos.

3. Métricas de evaluación de la robustez del sistema ante fallas en cascada En

CPPS, cualquier enlace defectuoso tiene el potencial de propagarse a través de relaciones de acoplamiento y evolucionar hacia fallas en cascada. Para analizar el impacto de las fallas iniciales en el sistema, definimos métricas de evaluación para ambas redes acopladas en función de la integridad de la topología de la red y las características operativas.

3.1. Métricas de la red de comunicación.

En la red de comunicaciones, empleamos tasas de supervivencia ajustadas de nodos/enlaces para evaluar la integridad de la topología de la red de comunicaciones. Teniendo en cuenta que los enlaces de comunicación en un estado sobrecargado tienen un flujo de información que excede su capacidad, contrariamente a las condiciones normales, estos enlaces funcionan de manera ineficiente y tienen una cierta probabilidad de pasar a un estado fallido. Por lo tanto, para distinguir entre enlaces normales y sobrecargados y representar con mayor precisión el impacto del estado de sobrecarga en el sistema, se utiliza el tamaño relativo del componente conectado más grande Gc para evaluar los enlaces de comunicación después del ajuste [29]. La tasa de supervivencia de nodos/enlaces después de una falla se utiliza para evaluar la integridad topológica, como se detalla a continuación:

$$Gc = \frac{\sum h \quad \Psi sh}{} \tag{13}$$

donde Nc es el número de nodos en la capa de acceso de la red de comunicaciones. Ψ denota el conjunto de nodos que no fallaron. Cuando los enlaces de información están en estado normal, sh = 1.

Sin embargo, cuando los enlaces de información están en condiciones de sobrecarga, sh se calcula de la siguiente manera:

$$= \delta Ch \quad \frac{\delta Ch - Lh\ sh}{- Ch} \tag{14}$$

Entonces, la tasa de supervivencia ajustada del nodo/enlace es

$$Fc = \frac{1}{2} \left(\frac{V_C'}{vc} + Gc \right) \tag{15}$$

Debido a la desconexión de los enlaces de información, la ruta desde los nodos de la capa de acceso de la red de comunicación al centro de control puede cambiar, lo que resulta en retrasos en la comunicación. En este artículo, el retraso de la comunicación de datos se simplifica y se calcula de la siguiente manera [32]: la transmisión de datos en la red de comunicación sigue el principio del camino más corto, y cada vez que pasa a través de un nodo de datos, el retraso aumenta en una unidad de tiempo τ. La unidad de tiempo de retraso refleja el retraso causado por la transmisión y el procesamiento de datos desde el nodo de origen al nodo de destino en la red de comunicación, incluido el paso a través de cada nodo de información y la ruta de comunicación al siguiente nodo de información.
Por lo tanto, el incremento del retardo de transmisión T causado por la red de comunicación se calcula de la siguiente manera:

$$T = \sum_{k\ Lc} T_{\pi c}' + \sum_{k\ Lc} t\pi c \tag{dieciséis}$$

donde $T_{\pi c}'$ y Tπc son el retardo de la ruta de transmisión de los mismos pares fuente-destino después y antes del fallo en cascada. Lc denota el conjunto de rutas de transmisión más cortas para todos los pares fuente-destino en la red de comunicación.

3.2. Métricas de la red eléctrica

En la red eléctrica física, utilizamos el impacto de falla Fp para evaluar la influencia de la falla en cascada [12], expresada como

$$Fp = 2^{\frac{1}{-}} \left(\frac{vp'}{\_\_\_} \right) + \frac{mi_p'}{Ep.} \tag{17}$$

donde vp' y Ep', respectivamente, son el número de nodos y enlaces fallidos en la red eléctrica. Vp y Ep, respectivamente, son el número total de nodos y enlaces en la red eléctrica física.

4. Estudio de caso y discusión

En esta sección, tomando como ejemplo el sistema de bus IEEE 39, la topología de la red eléctrica se muestra en la Figura 5 y la topología original se divide en cuatro regiones según principios de zonificación. De manera correspondiente, la red de comunicación de energía, generada de acuerdo con las reglas antes mencionadas, se representa en la Figura 6. En esta red, la capa de acceso consta de 39 nodos de comunicación, cada uno correspondiente a uno de los 39 nodos de energía , que se utilizan para cargar información de fallas. y emitir instrucciones de despacho.
La capa principal incluye cuatro nodos de control, cada uno de los cuales corresponde a su respectiva región de capa de acceso. El centro de despacho está equipado con dos nodos de control.

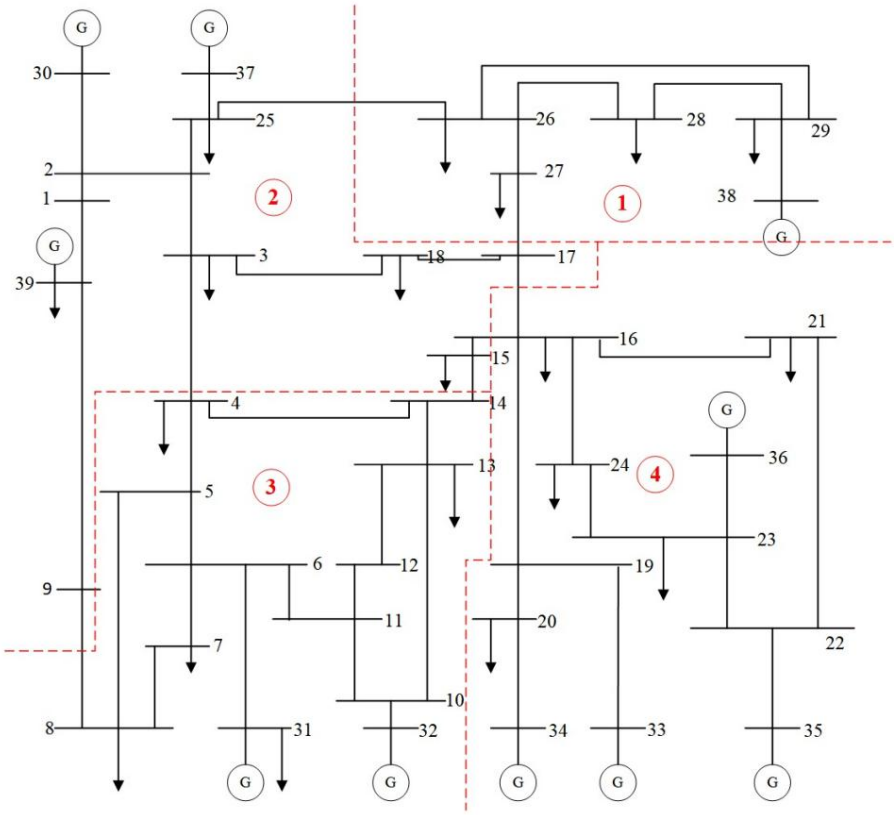


Figura 5. Diagrama de partición del sistema de bus IEEE 39.

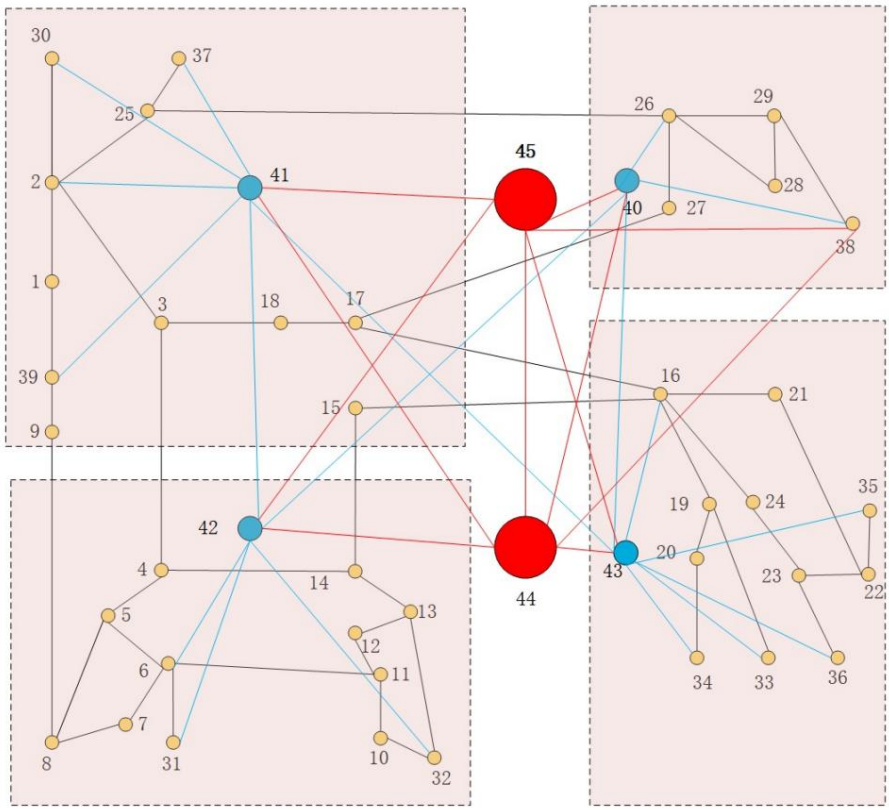


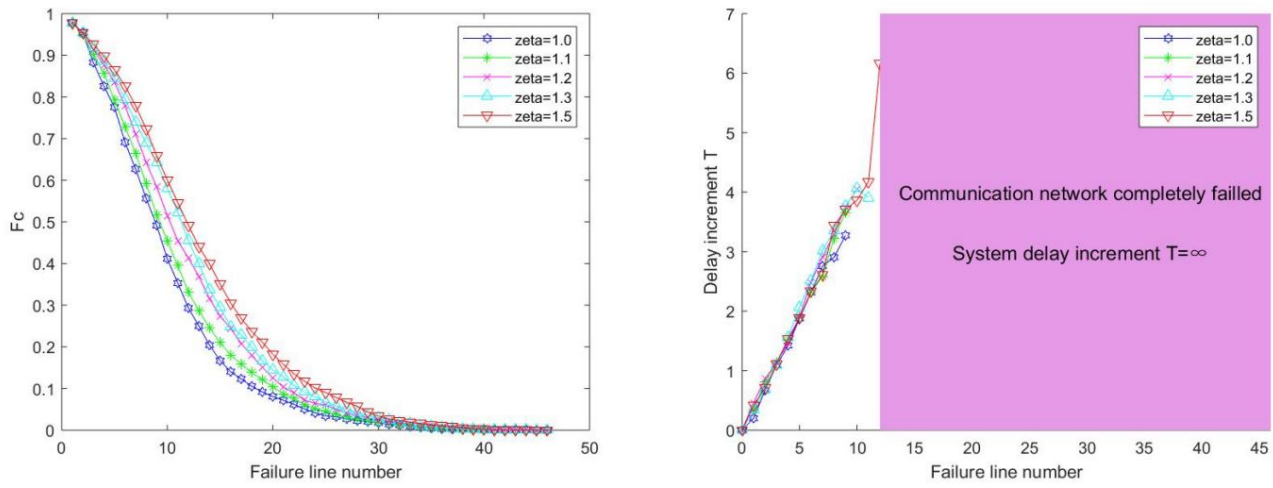
Figura 6. Diagrama de topología de la red de información generada por el sistema de 39 buses IEEE.

4.1. Impacto del coeficiente de sobrecarga

Para explorar el impacto del coeficiente de sobrecarga en la robustez de la red, se varió el número de fallas de enlace iniciales, y cada conjunto de enlaces defectuosos se generó aleatoriamente. Las simulaciones se ejecutaron de forma independiente 50 veces y el valor promedio se tomó como métrica de la red de comunicación. Para mayor claridad de presentación en los gráficos, se utilizaron descripciones textuales cuando el sistema de comunicación falló por completo, y para la explicación textual se seleccionó el conjunto de datos con el mayor número de enlaces fallidos que causaron la falla total del sistema de comunicación . Según la Figura 7a, a medida que aumenta el número de enlaces defectuosos iniciales, la tasa de supervivencia de los nodos/enlaces en la red de comunicación disminuye gradualmente. Cuando no se considera la capacidad de sobrecarga de los enlaces, la tasa de supervivencia de los nodos/enlaces en la red es la más baja. Cuando el coeficiente de sobrecarga es 1,2, la tasa de supervivencia de los nodos/enlaces en la red de comunicación mejora significativamente. Sin embargo, se puede observar que cuando los coeficientes de sobrecarga son 1,3 y 1,5, la tasa de supervivencia de los nodos/enlaces no mejora significativamente.

La Figura 7b muestra que el CPPS bajo diferentes coeficientes de sobrecarga exhibe una transición de percolación de primer orden, con una tendencia general similar. A medida que aumenta el número de líneas fallidas , la topología de la capa de información se altera y algunos nodos pierden sus rutas de transmisión de datos. Cambiar la ruta de transmisión aumenta el retraso del sistema. Cuando el coeficiente de sobrecarga es 1,5, la cantidad de enlaces fallidos que la red de comunicación puede soportar antes de colapsar por completo es la más alta. Como se ve en la Figura 7b y la Tabla 1, existe un umbral para el número inicial de enlaces fallidos, cerca del cual el efecto de falla en cascada se expande a todos los enlaces de comunicación, sin dejar ninguna ruta de transmisión entre los nodos de comunicación y el centro de control. En este punto, el sistema de comunicación falla por completo, imposibilitando el control de la red eléctrica. La Tabla 1 muestra los umbrales específicos para diferentes coeficientes de sobrecarga.

Por lo tanto, considerando los costos de construcción de la red, el coeficiente de sobrecarga se establece en 1,3 según la Figura 7 y la Tabla 1.



(a) (b)

Figura 7. La robustez de la red de comunicación y el incremento en el retraso de la comunicación bajo diferentes coeficientes de sobrecarga. (a,b) muestran respectivamente la relación entre la robustez de la red de comunicación y el retraso del sistema con el aumento en el número de fallas del enlace de comunicación bajo diferentes coeficientes de sobrecarga, mientras que otros parámetros permanecen constantes.

Tabla 1. Los umbrales del sistema bajo diferentes coeficientes de sobrecarga.

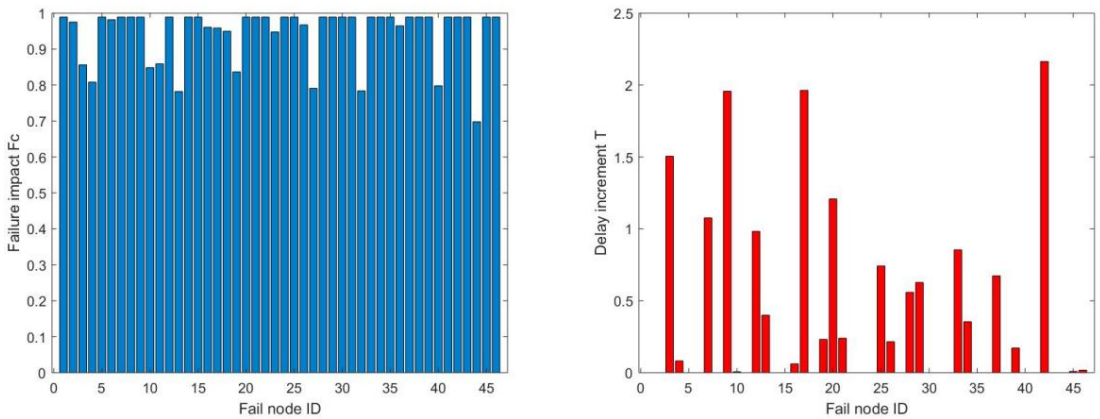
Coeficiente de sobrecarga	El umbral de las líneas fallidas iniciales	Umbral Porcentaje %
1,0	9	19.56
1,1	9	19.56
1,2	10	21,74
1,3	11	23,91
1,5	12	26.09

4.2. Impacto de los enlaces fallidos

En este apartado analizaremos el impacto en la transmisión de datos dentro del sector energético y redes de comunicación eliminando secuencialmente cada enlace de red de comunicación.

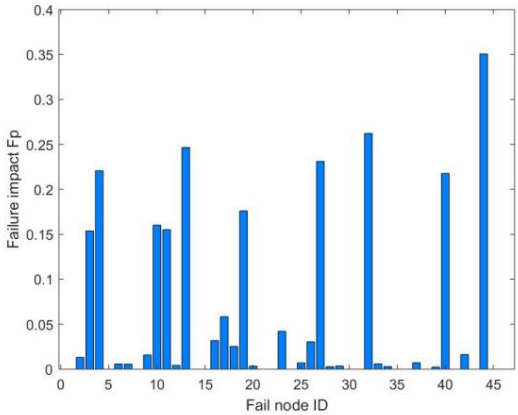
Como se muestra en la Figura 8, la mayoría de los enlaces de comunicación defectuosos tienen impactos variables sobre la integridad topológica y las características operativas del CPPS. Enlaces que causan Las interrupciones significativas del sistema se identifican como enlaces críticos. Bajo condiciones de falla N-1, todos Se enumeran posibles escenarios de ataques al sistema.

Como se muestra en la Figura 8, se puede observar que el fallo de una sola comunicación enlace puede reducir la tasa de supervivencia de nodos y enlaces en la red de comunicación al mínimo como 69,69%, induce un incremento de retraso del sistema de 2,1635 y causa una parálisis del 35,07% en los nodos y enlaces de la red eléctrica. También se observa que algunas fallas en los enlaces no afectar al resto de redes de comunicación y energía más allá del enlace defectuoso. Esto es porque, en fallas en cascada, la red de comunicación considera el estado de congestión de los enlaces, redistribuyendo el tráfico en enlaces fallidos y sobrecargados. Además, sobrecargado Los enlaces pueden mantener la operación durante un corto período antes de fallar, mejorando así la capacidad del sistema. robustez hasta cierto punto. Además, el vínculo de comunicación que provoca la máxima El incremento del retardo puede no necesariamente dar como resultado la tasa de supervivencia de nodo/enlace más baja en la red de comunicación o la tasa de falla más alta en los nodos/enlaces de la red eléctrica. Sin embargo, Las tasas de falla de nodos/enlaces en las redes de comunicación y energía son notablemente similar. Por ejemplo, cuando falla el enlace 42, se produce el incremento de retraso más alto de 2,1635. En En este punto, la tasa de supervivencia de los nodos/enlaces de comunicación es del 98,91%, y la tasa de fallo de nodos/enlaces de potencia es del 3,46%. Esto se debe a que la falla del enlace 42 cambia la ruta de nodo de comunicación 27 al centro de control, desde la ruta original 27→26→40→44 hasta 27→17→16→43→44, lo que provoca un retraso en la comunicación. La tasa de fracaso de la comunicación. el nodo 27 es 36,06%. Sin embargo, eliminar el enlace 42 no causa falla o sobrecarga en otros enlaces, por lo que hay 45 enlaces operativos y el número de nodos de comunicación efectivos permanece sin cambios en comparación con antes de la falla. Debido a la relación de acoplamiento, el La tasa de falla del nodo de energía 27 también es del 36,06%, lo que en última instancia conduce a la falla del suministro de energía. nodo 27.



(a) (b)

Figura 8. Continuación.



(C)

Figura 8. El impacto de las fallas iniciales en el sistema ciberfísico. (a,b) muestran el impacto de cada falla del enlace de comunicación en la robustez y el incremento de retardo del sistema de comunicación, mientras que (c) muestra el impacto de cada falla del enlace de comunicación en la red eléctrica física.

4.3. Impacto de las estrategias de ataque

En el apartado anterior se analizó el impacto N-1 provocado por el fallo de un único enlace de comunicación. A medida que aumenta el número de líneas defectuosas, diferentes conjuntos iniciales de líneas defectuosas tendrán diferentes efectos en el sistema. Por lo tanto, basándonos en las características estructurales y eléctricas , empleamos tanto un ataque de línea aleatorio como un ataque de línea crítica para eliminar los enlaces de la red de comunicación uno por uno, analizando el impacto de diferentes estrategias de ataque en la transmisión de datos de la red de comunicación y energía .

La estrategia de ataque de línea crítica se deriva de la clasificación inicial de importancia de centralidad de intermediación de poder de la red eléctrica. Debido a la incertidumbre de la estrategia de ataque de línea aleatoria, cada línea defectuosa en la estrategia de ataque aleatorio se genera aleatoriamente y los resultados se promedian después de 50 ejecuciones independientes.

Como se muestra en la Figura 9, en comparación con los ataques aleatorios, el CPPS exhibe una alta vulnerabilidad ante ataques en líneas críticas. Bajo ataques a líneas críticas, incluso considerando previamente la redistribución del flujo de información debido a la sobrecarga de enlaces, la red todavía sufre daños significativos. Esto se debe a que la falla de las líneas críticas interrumpe las rutas que transmiten información al centro de control, impidiendo la carga de información sobre fallas de energía . En consecuencia, el centro de control no puede responder y los nodos fuente que transmiten información en la red de comunicación provocarán la falla de los nodos de energía acoplados, expandiendo así la escala de fallas y acelerando la propagación de fallas en cascada y el colapso del sistema.

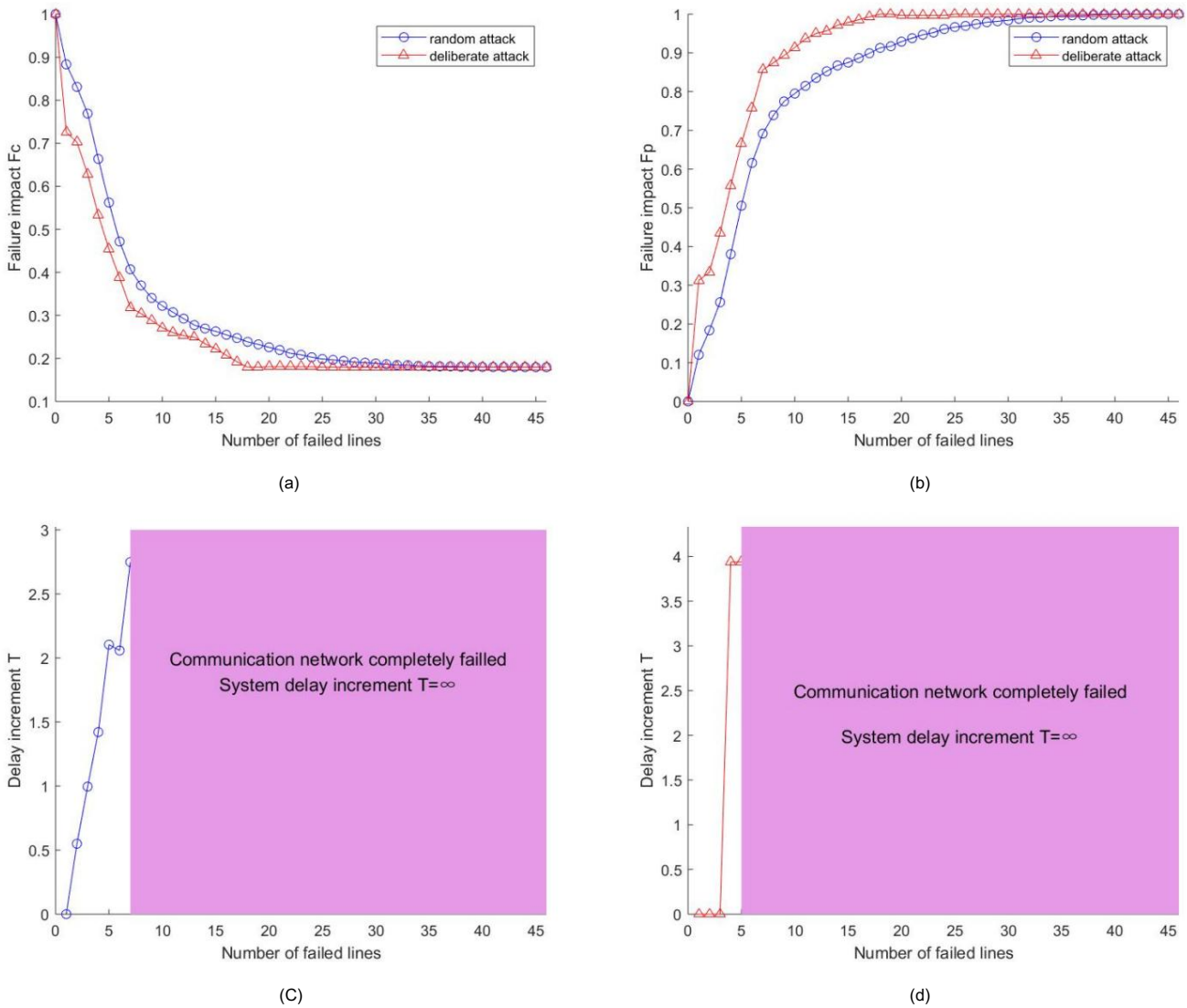


Figura 9. Impacto de las estrategias de ataque en el sistema ciberfísico. (a – d) ilustran la relación entre las métricas de evaluación de la red acoplada y el número de enlaces de comunicación fallidos bajo dos estrategias de ataque diferentes (ataque aleatorio y ataque deliberado). Específicamente, (a) muestra la relación entre la tasa de supervivencia de nodos/enlaces en la red de comunicación eléctrica y el número de enlaces de comunicación fallidos, (b) muestra la relación entre la tasa de falla de nodos/enlaces en la red de comunicación eléctrica y la Número de enlaces de comunicación fallidos . (c,d) representan respectivamente la tendencia del incremento del retraso en el sistema bajo ataque aleatorio y ataque deliberado.

5. Conclusiones

Este artículo investiga la propagación dinámica de fallas en cascada en CPPS. Primero, se analiza la relación de acoplamiento entre la red de comunicación y la red eléctrica física , y se genera la estructura topológica del sistema de comunicación correspondiente en función de la red eléctrica. A continuación, se establece un modelo de distribución de sobrecarga para el flujo de información dentro de la red de comunicación para abordar las sobrecargas resultantes de fallas. En la red eléctrica, se establece un modelo de distribución de capacidad de carga basado en la intermediación del flujo de potencia para manejar la complejidad de los cálculos del flujo de potencia debido a fallas. Para describir con mayor precisión las fallas de nodo/línea causadas por transferencias de flujo de energía debido a fallas, se mejora la teoría de percolación tradicional para establecer un modelo de propagación de fallas para redes acopladas.

En general, este artículo considera de manera integral tanto las características eléctricas de La red eléctrica física y las características de transmisión de información de la red eléctrica. red de comunicación al establecer el modelo de acoplamiento. Mejora el acoplado modelo de red tanto desde la estructura topológica como desde las características funcionales y construye un modelo de fallas en cascada que refleja con mayor precisión escenarios del mundo real. La mayoría de los estudios existentes se basan en la teoría de la percolación y sólo analizan el fallo en cascada. proceso de CPPS desde la perspectiva de la integridad topológica de la red. Este enfoque falla capturar de manera realista los fenómenos de isla en las operaciones del sistema eléctrico, por lo que es necesario una mejora de la teoría de la percolación.

Las simulaciones analizan el impacto de las líneas defectuosas en la capacidad de sobrecarga de los enlaces de las redes de comunicación, la transmisión de datos, la capacidad de supervivencia de la red y la energía física. Tasas de fallas en la red. Los resultados indican que incluso un pequeño aumento en la capacidad del enlace puede mejorar significativamente la robustez de la red. Además, el rendimiento del sistema se degrada. cuando se atacan enlaces en posiciones topológicas críticas en la red de comunicación, lo que lleva a mayores incrementos de retraso del sistema en comparación con los ataques aleatorios y provoca la que la red de la capa de información colapse antes.

En resumen, estos hallazgos ofrecen información valiosa para la planificación futura de redes eléctricas. Sin embargo, estos estudios se basan en condiciones posteriores a la falla, considerando solo las interacciones entre nodos como una falla normal o total para lograr el despacho del flujo de energía por parte del centro de control. En realidad, el proceso de interacción entre nodos de poder e información. los nodos es extremadamente complejo y a menudo sólo se analiza cualitativamente. Además, en comparación Para el despacho del flujo de energía, programar el flujo de información es relativamente simple. Reduciendo el La aparición de fallas desde la fuente del lado cibernético es una solución más eficiente y confiable. en lugar de establecer un mecanismo de resistencia a fallas en cascada en el lado de la red eléctrica. Por lo tanto, nuestra investigación futura se centrará en asignar flujos de información críticos a fuentes confiables. rutas en condiciones de sobrecarga del enlace de comunicación para garantizar la accesibilidad y reducir la probabilidad de fallos en cascada causados por interrupciones en la comunicación.

Contribuciones de los autores: Conceptualización, XL y YL; Metodología, XL y YL; Escritura—original borrador, XL; Redacción: revisión y edición, YL y TX Todos los autores han leído y aceptado las versión publicada del manuscrito.

Financiamiento: Este trabajo fue apoyado por la Fundación Nacional de Ciencias Naturales de China bajo Beca 62062068 y Programa de Jóvenes Líderes Académicos y Técnicos de la Provincia de Yunnan bajo Subvención 202305AC160077.

Declaración de disponibilidad de datos: Las contribuciones originales presentadas en el estudio se incluyen en el artículo, cualquier consulta adicional puede dirigirse al autor correspondiente.

Agradecimientos: Gracias por la ayuda del Kunming AI Computing Center.

Conflictos de intereses: Los autores declaran no tener conflictos de intereses.

Abreviaturas y nomenclatura

CPPS	Sistema de energía ciberfísico;
Fiyi	Demanda de flujo de información en el enlace eij;
qij	La relación entre la capacidad del enlace y la carga del enlace;
ρ	El umbral de falla del enlace;
——	Los conjuntos de nodos de potencia;
Ep.	Los conjuntos de líneas de transmisión de energía eléctrica;
q	El valor de modularidad de la unión combinada de dos comunidades;
qi	La proporción de la producción del generador i con respecto a la producción total del sistema;
δ	La capacidad del enlace para manejar flujo de información adicional;
ki	Los grados del nodo i;
θ	El parámetro que ajusta el flujo de información;
Ic,ij	La cantidad de información transmitida por el enlace eij;
CC,ij	La capacidad del enlace eij;

$C_{c,ij}^{máximo}$	El caudal máximo que puede soportar el enlace eij ;
wc,ij	El peso del enlace eij;
α	El coeficiente de capacidad;
β	El coeficiente de capacidad;
min(Sm, Sn)	El peso de la intermediación de flujo de una sola línea;
Pij,m	La porción del flujo de energía en la línea eij que se origina en el generador m;
Pij,n	La porción del flujo de potencia en la línea eij dirigida hacia la carga n;
Pij	La potencia activa a través de la línea eij;
PLn	La carga activa en el nodo de carga n;
A_{unm}^{-1}	Los elementos de la matriz de distribución de orden inverso;
PGM	La salida activa del nodo generador m;
FBij	La intermediación del flujo del borde eij en la red eléctrica;
Lp(ij)	La carga de energía en el borde eij de la red eléctrica;
Cp(ij)	La capacidad del borde eij en la red eléctrica;
γ	El parámetro de tolerancia;
Δ_{ak}	La estrategia de distribución para la sucursal sobrecargada;
fc	La tasa de supervivencia ajustada del nodo/enlace después de una falla en la red de comunicación;
t	El incremento del retardo de transmisión;
fp	La tasa de nodo/enlace fallido después de una falla en la red eléctrica.

Referencias

1. Abdelmalak, M.; Venkataramanan, V.; Macwan, R. Un estudio de los métodos de modelado de sistemas de energía ciberfísicos para la energía futura. sistemas. Acceso IEEE 2022, 10, 99875–99896. [Referencia cruzada]
2. Li, Y.; Wang, B.; Wang, H.; Mamá, F.; Mamá, H.; Zhang, J.; Zhang, Y.; Mohamed, MA Un interdependiente eficaz de nodo a borde Análisis de redes y vulnerabilidades para redes eléctricas acopladas digitales. En t. Trans. eléctrico. Sistema de energía. 2022, 2022, 5820126. [Referencia cruzada]
3. Gao, X.; Peng, M.; Chi, KT; Zhang, H. Un modelo estocástico de dinámica de fallas en cascada en sistemas de energía ciberfísicos. IEEE Sistema. J. 2020, 14, 4626–4637. [Referencia cruzada]
4. Liang, G.; Weller, SR; Zhao, J.; Luo, F.; Dong, ZY El apagón de Ucrania de 2015: implicaciones de los ataques de inyección de datos falsos. IEEE Trans. Sistema de energía. 2016, 32, 3317–3318. [Referencia cruzada]
5. Zhao, Q.; Qi, X.; Hua, M.; Liu, J.; Tian, H. Reseña de los recientes apagones y la Ilustración. En Actas del CIRED 2020 Taller de Berlín (CIRED 2020), conferencia en línea, 22 y 23 de septiembre de 2020; IET: Stevenage, Reino Unido, 2020; Volumen 2020; págs. 312–314.
6. Wu, J.; Chen, Z.; Zhang, Y.; Xia, Y.; Chen, X. Recuperación secuencial de redes complejas que sufren apagones por fallas en cascada. Traducción IEEE. Neto. Ciencia. Ing. 2020, 7, 2997–3007. [Referencia cruzada]
7. Peng, C.; Wu, J.; Tian, E. Control H∞ desencadenado por eventos estocásticos para sistemas en red bajo ataques de denegación de servicio. IEEE Trans. Sistema. Hombre Cibernético. Sistema. 2021, 52, 4200–4210. [Referencia cruzada]
8. Hu, S.; Yue, D.; Han, QL; Xie, X.; Chen, X.; Dou, C. Control activado por eventos basado en observadores para sistemas lineales en red Sujeto a ataques de denegación de servicio. Traducción IEEE. Cibern. 2020, 50, 1952–1964. [Referencia cruzada]
9. Buldyrev, SV; Parshani, R.; Pablo, G.; Stanley, ÉL; Havlin, S. Cascada catastrófica de fallas en redes interdependientes. Naturaleza 2010, 464, 1025–1028. [Referencia cruzada]
10. Huang, X.; Gao, J.; Buldyrev, SV; Havlin, S.; Stanley, HE Robustez de las redes interdependientes bajo ataque dirigido. Física. Rev. E—Stat. Física de materia blanda no lineal. 2011, 83, 065101. [Referencia cruzada] [PubMed]
11. Cao, YY; Liu, RR; Jia, CX; Wang, BH Percolación en redes complejas multicapa con conectividad e interdependencia estructuras topológicas. Comunitario. Ciencia no lineal. Número. Simultáneo. 2021, 92, 105492. [Referencia cruzada]
12. Chen, L.; Yue, D.; Dou, C.; Cheng, Z.; Chen, J. Robustez de los sistemas de energía ciberfísicos en fallas en cascada: supervivencia de grupos interdependientes. En t. J. Electr. Sistema de energía eléctrica. 2020, 114, 105374. [Referencia cruzada]
13. Pan, H.; Li, X.; Na, C.; Cao, R. Modelado y análisis de fallas en cascada en sistemas de energía ciberfísicos bajo diferentes estrategias de acoplamiento. Acceso IEEE 2022, 10, 108684–108696. [Referencia cruzada]
14. Jianfeng, D.; Jian, Q.; Jing, W.; Xuesong, W. Un método de evaluación de la vulnerabilidad del sistema de energía ciberfísico considerando Fallo de las infraestructuras de la red eléctrica. En actas de la Conferencia IEEE sobre energía y energía sostenible (iSPEC) de 2019, Beijing, China, 21 a 23 de noviembre de 2019; IEEE: Piscataway, Nueva Jersey, EE. UU., 2019; págs. 1492-1496.
15. Tian, M.; Dong, Z.; Gong, L.; Wang, X. Estrategias de refuerzo de línea para sistemas de energía resilientes considerando la topología cibernética interdependencia. Confiable. Ing. Sistema. Seguro. 2024, 241, 109644. [Referencia cruzada]
16. Chen, L.; Yue, D.; Dou, C.; Xie, X.; Li, S.; Zhao, N.; Zhang, T. Impacto de la falla en cascada en la distribución de energía y los datos transmisión en sistemas de energía ciberfísicos. Traducción IEEE. Neto. Ciencia. Ing. 2023, 11, 1580–1590. [Referencia cruzada]
17. Zhong, J.; Zhang, F.; Yang, S.; Li, D. Restauración de la red interdependiente contra fallas por sobrecarga en cascada. Física. Una estadística. Mec. Su aplicación. 2019, 514, 884–891. [Referencia cruzada]

18. Ding, D.; Wu, H.; Yu, X.; Wang, H.; Yang, L.; Wang, H.; Kong, X.; Liu, Q.; Lu, Z. Evaluación de vulnerabilidad del sistema de energía ciberfísica basada en un modelo mejorado de fallas en cascada. J. Electr. Ing. Tecnología. 2024, 1–12. [Referencia cruzada]
19. Wang, S.; Gu, X.; Chen, J.; Chen, C.; Huang, X. Estrategia de mejora de la robustez de los sistemas ciberfísicos con débiles interdependencia. Confiable. Ing. Sistema. Seguro. 2023, 229, 108837. [Referencia cruzada]
20. Ghasemi, A.; de Meer, H. Robustez de las redes eléctricas y de comunicación interdependientes ante fallas en cascada. Traducción IEEE . Neto. Ciencia. Ing. 2023, 10, 1919-1930. [Referencia cruzada]
21. Cao, R.; Dong, X.; Wang, B.; Liu, K. Discusión sobre protección y cortes en cascada desde el punto de vista de la comunicación. En Actas de la Conferencia Internacional de 2011 sobre Automatización y Protección Avanzada de Sistemas Eléctricos, Beijing, China, 16 a 20 de octubre de 2011; IEEE: Piscataway, Nueva Jersey, EE. UU., 2011; Volumen 3, págs. 2430–2437.
22. Zhang, G.; Shi, J.; Huang, S.; Wang, J.; Jiang, H. Un modelo de falla en cascada considerando las características de operación del capa de comunicación. Acceso IEEE 2021, 9, 9493–9504. [Referencia cruzada]
23. Gao, X.; Peng, M.; Chi, KT Análisis de robustez de sistemas de energía ciberacoplados con consideraciones de interdependencia de estructuras, operaciones y comportamientos dinámicos. Física. Una estadística. Mec. Su aplicación. 2022, 596, 127215. [Referencia cruzada]
24. Wang, F.; Li, D.; Xu, X.; Wu, R.; Havlin, S. Propiedades de percolación en un modelo de tráfico. Eurofis. Letón. 2015, 112, 38001. [Referencia cruzada]
25. Saberi, M.; Hamedmoghadam, H.; Ashfaq, M.; Hosseini, SA; Gu, Z.; Shafiei, S.; Nair, DJ; Dixit, V.; Gardner, L.; Waller, ST; et al. Un simple proceso de contagio describe la extensión de los atascos en las redes urbanas. Nat. Comunitario. 2020, 11, 1616. [Referencia cruzada]
26. Liu, W.; Liang, C.; Xu, P.; Dan, Y.; Wang, J.; Wang, W. Identificación de línea crítica en sistemas de energía basados en flujo intermedio. Proc. CSEE 2013, 33, 90–98.
27. Wang, T.; Sol, C.; Gu, X.; Qin, X. Modelado y análisis de vulnerabilidad de redes acopladas de comunicación de energía eléctrica. Proc. CSEE 2018, 38, 3556–3567.
28. Newman, ME Encontrar estructura comunitaria en redes utilizando vectores propios de matrices. Física. Rev. E—Stat. Suave no lineal Materia Física. 2006, 74, 036104. [Referencia cruzada]
29. Chen, CY; Zhao, Y.; Gao, J.; Stanley, HE Modelo no lineal de falla en cascada en redes complejas ponderadas considerando bordes sobrecargados. Ciencia. Rep. 2020, 10, 13428. [CrossRef]
30. Stauffer, D.; Aharony, A. Introducción a la teoría de la percolación; Taylor y Francis: Abingdon, Reino Unido, 2018.
31. Motter, AE; Lai, YC Ataques basados en cascada en redes complejas. Física. Rev.E 2002, 66, 065102. [CrossRef]
32. Han, Y.; Guo, C.; Zhu, B.; Xu, L. Modele fallas en cascada en un sistema de energía ciberfísico basado en una teoría de percolación mejorada. Automático. eléctrico. Sistema de energía. 2016, 40, 30–37.

electrónica

MDPI

Artículo

Una metodología práctica de evaluación de la seguridad para sistemas eléctricos Operaciones considerando la incertidumbre

Nhi Thi Ai Nguyen [1], Dinh Duong Le 1,*, Van Duong Ngo [1], Van Kien Pham [1] y Van Ky Huynh [2]

[1] Facultad de Ingeniería Eléctrica, Universidad de Danang—Universidad de Ciencia y Tecnología, Danang 550000, Vietnam; ntanhi@dut.udn.vn (NTAN); nvduong@dut.udn.vn (VDN); pvkien@dut.udn.vn (VKP)
[2] La Universidad de Danang, Danang 550000, Vietnam; hvky@ac.udn.vn
* Correspondencia: lddduong@dut.udn.vn

Resumen: Hoy en día, las fuentes de energía renovables (FER) se están integrando cada vez más en los sistemas eléctricos. Esto significa agregar más fuentes de incertidumbre al sistema eléctrico. Para abordar la incertidumbre de las variables aleatorias de entrada (RV) en los problemas de cálculo y análisis de sistemas de energía, se han introducido técnicas de flujo de energía probabilístico (PPF) que han demostrado ser efectivas. Actualmente , aunque existen muchas técnicas propuestas para resolver el problema de PPF, el método de simulación de Monte Carlo (MCS) todavía se considera el método con mayor precisión y sus resultados se utilizan como referencia para la evaluación de otros métodos. Sin embargo, el MCS a menudo requiere una intensidad computacional muy alta, lo que dificulta la aplicación práctica, especialmente en sistemas de energía a gran escala. En el artículo actual, se propone una técnica avanzada de agrupación de datos para procesar datos de entrada de RV con el fin de disminuir la carga computacional de resolver el problema de PPF y al mismo tiempo mantener un nivel aceptable de precisión. El método propuesto se puede aplicar eficazmente para resolver problemas prácticos en el horizonte temporal de operación de los sistemas de energía. El enfoque desarrollado se prueba en el sistema de bus IEEE-300 modificado, lo que indica un buen rendimiento en la reducción del t

Palabras clave: flujo de potencia probabilístico; Sistema de poder; energía renovable

Cita: Nguyen, NTA; Le, DD; Ngo, VD; Pham, VK; Huynh, VK Una metodología práctica de evaluación de la seguridad para las operaciones de sistemas eléctricos considerando la incertidumbre. Electrónica 2024, 13 , 3068. https://doi.org/10.3390/electronics13153068

Editor académico: Ahmed Abu-Siada

Recibido: 26 de junio de 2024
Revisado: 31 de julio de 2024
Aceptado: 31 de julio de 2024
Publicado: 2 de agosto de 2024

1. Introducción

Durante la operación de un sistema de energía, los parámetros del modo de operación, como el voltaje en los autobuses, la potencia transmitida a través de las ramas, etc., deben calcularse periódicamente para evaluar la seguridad del sistema comparando los parámetros con sus límites permitidos. En caso de existir un riesgo para la seguridad, se deben proponer soluciones razonables para resolverlo. El flujo de energía determinista (DPF) es una de las herramientas esenciales para la operación y planificación del sistema eléctrico. Sin embargo, durante el proceso de cálculo, el enfoque tradicional utiliza valores fijos de inyecciones de potencia nodal (de generación de energía, carga, etc.) y la estructura de red conocida para que no se consideren las fuentes de incertidumbre de estos factores. Esta es la principal limitación del método tradicional de flujo de potencia (PF) [1].

Para superar la desventaja mencionada anteriormente, se propuso el PPF y se ha convertido en una herramienta de cálculo muy eficaz. La carga, la generación de energía de una planta y el funcionamiento de un elemento como una línea, un transformador, etc., pueden seguir ciertas reglas de probabilidad . En particular, para los sistemas energéticos actuales, cuando se conectan al sistema fuentes de energía renovables adicionales, como la solar y la eólica, etc., modelar la intermitencia de las energías renovables es un gran desafío. La intermitencia suele cambiar muy rápida y estocásticamente, aumentando el nivel de incertidumbre en el sistema. Por tanto, se requiere un método de cálculo para poder integrar las incertidumbres en el proceso de cálculo. Al utilizar métodos PPF, las salidas, es decir, el voltaje en las barras, el PF transmitido en las ramas, etc., también cambian aleatoriamente según una ley de distribución de probabilidad [1]. El análisis PPF nos permite valorar la probabilidad de sobrecarga de la línea, la probabilidad de sobre/subtensión, etc. A partir de ahí, depe

características del sistema y la gravedad de la infracción, el operador podría considerar y sugerir soluciones apropiadas para mejorar la seguridad del sistema.

El enfoque PPF fue introducido por primera vez por Borkowska en 1974 [2] y, desde entonces, se han propuesto varios trabajos de investigación para la PPF en todo el mundo. Generalmente, los métodos para calcular el FP utilizando la técnica PPF se pueden clasificar en tres categorías principales, es decir, enfoques numéricos, analíticos y de aproximación.

El enfoque analítico [3–6] hace uso de algoritmos y técnicas analíticas como las técnicas de convolución y acumulativa. La aplicación de estas técnicas analíticas combinadas con la relación de la entrada y salida de un problema PPF permite determinar la función de distribución de los RV de salida, como la transmisión de potencia en la línea, voltaje nodal, ángulo de fase, etc., según el sistema. parámetros, por ejemplo, la impedancia total de la línea, la impedancia total del transformador, etc., y distribuciones de probabilidad de los RV de entrada de la carga y generación de energía de los generadores tradicionales y RES, así como el estado operativo de los dispositivos. La relación entre la entrada y la salida del problema de cálculo de PF no es lineal. Sin embargo, el método analítico funciona bien con una relación lineal entre la entrada y la salida del problema. Por lo tanto, en primer lugar, es necesario linealizar la relación utilizando una técnica de expansión, por ejemplo, la expansión de Taylor. Una de las ventajas destacadas del enfoque analítico es que puede dar resultados muy rápidos.

Entre los enfoques acumulativo y convolucional, el enfoque de convolución es más intensivo desde el punto de vista computacional que el cumulante. Por lo tanto, actualmente, el enfoque acumulativo es más popular que el enfoque convolucional. Para lograr las funciones de distribución para los RV de salida, el enfoque acumulativo a menudo se usa simultáneamente con técnicas de expansión como la expansión de Gram-Charlier o Cornish-Fisher [4]. Debido a la ventaja de un cálculo rápido, el enfoque analítico podría aplicarse en la práctica a un sistema eléctrico a gran escala. Sin embargo, el enfoque analítico tiene algunos inconvenientes. En primer lugar, la precisión del enfoque analítico se ve significativamente afectada por el uso de técnicas que linealizan la relación de entrada y salida, especialmente cuando el RV de entrada cambia en un amplio rango, por ejemplo, en el caso de las FER. En segundo lugar, el enfoque analítico utiliza técnicas de expansión que pueden funcionar bien en el caso de que las funciones de distribución de los RV de entrada sean de distribución gausiana o cercanas a la distribución gausiana. De hecho, las funciones de distribución de los RV de entrada del problema PPF para un sistema de potencia, en la práctica, a menudo siguen una distribución no gausiana, por lo que los resultados obtenidos serán limitados. Para poder integrar funciones de distribución discreta de los RV de entrada en el proceso de cálculo, se propone el método de Von Mises [1].

El enfoque típico para el grupo de aproximación en el cálculo de la FPP es el enfoque de estimación puntual [7,8]. En este enfoque, el RV de entrada se descompone en una secuencia de pares de valor y peso. A continuación, se calcula el momento del RV de salida como función del RV de entrada y luego se obtiene la función de distribución del RV de salida. El método de estimación puntual puede proporcionar resultados relativamente rápidos. Además, a diferencia del enfoque analítico, este enfoque utiliza la relación no lineal entre la entrada y la salida del problema de cálculo de PF. Sin embargo, la principal limitación del enfoque de estimación puntual es que su precisión disminuye a medida que aumenta el orden del momento. Otro inconveniente es que el tiempo de cálculo requerido aumenta significativamente a medida que aumenta el número de RV

Un enfoque numérico típico es MCS [9-14]. En MCS, se muestrean los RV de entrada y luego se realiza el cálculo del DPF para todas las muestras. Repite la simulación con una gran cantidad de muestras para obtener un resultado de alta precisión. MCS utiliza la relación no lineal entre la entrada y la salida del problema PF, como el enfoque tradicional. La principal ventaja del enfoque MCS es que proporciona resultados muy precisos y fiables. Además, las distribuciones de probabilidad de los RV de entrada en MCS son fáciles de representar. Sin embargo, el mayor inconveniente es que el volumen de cálculo es pesado y el tiempo de cálculo es relativamente largo, lo que dificulta la solicitud de una red eléctrica práctica a gran escala. Para reducir la carga computacional del método MCS, se proponen varios algoritmos de agrupamiento para reducir la cantidad de muestras y luego se ejecuta DPF para cada grupo en lugar de ejecutarlo para todas las muestras. En [15], se propone utilizar un algoritmo PSO para

la tarea de agrupación, mientras que K-means se utiliza en [16,17]. Cada algoritmo de agrupamiento tiene sus propias debilidades. Para PSO, su tasa de convergencia iterativa es modesta y, a menudo, se queda estancado en los óptimos locales al tratar con conjuntos de datos de alta dimensión. Para el algoritmo K-means, es sensible a la elección de k y es difícil encontrar el k óptimo para un conjunto de datos determinado. Además, K-means no se adapta bien a problemas grandes, razón por la cual, en el presente estudio, se centra en este problema.

Del análisis anterior, se puede ver que cada enfoque de FPP tiene sus propias características. Características, ventajas y desventajas.

Además de la descripción general anterior de la FPP, también se han encontrado avances recientes para resolver diversos problemas relacionados con la incertidumbre. En [18], se desarrolla un método de despacho coordinado robusto de dos etapas para microrredes multienergéticas para aliviar todos los efectos negativos de las diversas incertidumbres de la energía eólica y las cargas. El método de intervalo se emplea en [18] para caracterizar las incertidumbres. En [19] se propone una región de operación comprometida con las emisiones de carbono (CCEOR) de sistemas energéticos integrados (IES). El método desarrollado convierte el modelo CCEOR no lineal incierto propuesto en un modelo CCEOR determinista entero mixto convexo. En [20], para tener en cuenta diferentes tipos de incertidumbres derivadas de los resultados de desastres, eventos extremos, cargas y generación renovable, tanto las medidas de la etapa previa a la restauración como las de la etapa en tiempo real se coordinan mediante un método de programación estocástica de dos etapas .

Las principales contribuciones de este artículo se resumen a continuación: (1) El objetivo central de este estudio es desarrollar un enfoque para calcular la PPF que garantice un cierto nivel de precisión en comparación con MCS pero que debe dar resultados muy rápidos cercanos al “tiempo real”. ”Funcionamiento del sistema eléctrico. Para explotar las ventajas de la precisión del método MCS y al mismo tiempo reducir el tiempo y el volumen de cálculo, se propone una técnica de agrupamiento en tiempo real combinada con MCS en el cálculo de PPF. La técnica de agrupamiento aplicada es simple pero efectiva y adecuada para una aplicación práctica. La gran cantidad de muestras de RV de entrada del problema de cálculo de PPF utilizando MCS se reduce significativa y efectivamente, por lo que el cálculo de PPF proporciona resultados rápidos. Gracias a esta característica sobresaliente, el enfoque PPF se puede aplicar a grandes sistemas de energía en la práctica y al marco temporal operativo. (2) Además, en este artículo también se presenta en detalle la discusión sobre la aplicación de los métodos PPF en el cálculo y análisis de sistemas eléctricos. Esto proporciona una imagen más clara e intuitiva de la aplicación del análisis PPF tanto en la planificación como en la operación de sistemas eléctricos. También se señalan las limitaciones actuales de los métodos PPF en general para proporcionar temas para futuras investigaciones.

El resto del artículo se estructura de la siguiente manera. La Sección 2 presenta la metodología desarrollada, mientras que los resultados obtenidos mediante el enfoque desarrollado se discuten en la Sección 3. En la Sección 4, se brinda una discusión adicional sobre la aplicabilidad de varios métodos PPF en problemas de análisis de sistemas de energía. Las observaciones finales se proporcionan en la Sección 5.

2. Metodología 2.1.

Técnica de agrupación en clústeres en

tiempo real La agrupación en clústeres es la división de datos en varios grupos de modo que los puntos de datos del mismo grupo tengan características similares entre sí y sean diferentes a los puntos de datos de otros grupos. Básicamente es la tarea de dividir los datos de un conjunto de datos en función de las similitudes y diferencias entre ellos. Entre las técnicas de clustering, K-means es conocida como la más popular y se aplica en todos los campos debido a su simplicidad, eficiencia, escalabilidad y facilidad de implementación. Puede manejar grandes conjuntos de datos de forma eficaz, lo que lo convierte en una opción práctica para numerosas aplicaciones. En [21] se presenta una revisión exhaustiva de la aplicación de la agrupación de K-medias en los sistemas de energía modernos. El algoritmo de agrupamiento de K-medias es un tipo de aprendizaje automático no supervisado que divide el conjunto de datos sin etiquetar en k grupos diferentes mediante un algoritmo iterativo.

El algoritmo K-medias se implementa de la siguiente manera:

Paso 1: elija aleatoriamente k puntos o centroides de los datos considerados para inicializar los grupos o conglomerados;

Paso 2: Para cada punto del conjunto de datos, calcule la distancia entre el punto y cada de los k centroides; asigne cada punto a su centroide más cercano para formar k grupos;

Paso 3: Reemplace el centroide de cada grupo con la media de todos los puntos de datos asignados al cúmulo;

Paso 4: repita los pasos 2 y 3 hasta que los centroides ya no cambien significativamente o después de un número máximo de iteraciones preseleccionado. Los resultados obtenidos son los últimos centroides del grupo y los puntos de datos asignados a los grupos.

Aunque el método K-means tiene muchas ventajas, tiene algunas desventajas que pueden afectar su aplicabilidad y rendimiento. No se considera que el algoritmo K-means tenga buena escalabilidad para problemas grandes. Es sensible a la elección de k y es difícil encontrar el k óptimo para un conjunto de datos determinado. K-medias converge a un mínimo local, por lo que diferentes inicializaciones darán resultados diferentes.

K-means puede llevar mucho tiempo con conjuntos de datos grandes. Para un conjunto de datos que incluye n puntos de datos, es necesario ejecutarlo O(nkT) veces para calcular las distancias entre los n puntos de datos y cada uno de los k centroides (T es el número de iteraciones) [22]. En el algoritmo K-means, cada iteración toma un tiempo proporcional a k y n. Esto explica por qué el algoritmo K-means tiene poca escalabilidad. El tiempo de ejecución aumentará al aumentar n o k, o ambos. Por lo tanto, su eficiencia se puede mejorar significativamente disminuyendo el tiempo de ejecución relacionado con n. Para abordar el problema de la mala escalabilidad, en [22], los autores desarrollan un enfoque K-means-lite que puede obtener los centroides buscados en tiempo O(1) con respecto a n y exhibe un factor de aceleración mejorado. a medida que k y n aumentan. También se ha demostrado que la precisión aumenta. Se utiliza la técnica de inferencia estadística, en la que los k centroides se calculan utilizando unas pocas muestras pequeñas, en lugar de una comparación exhaustiva repetida entre centroides y puntos de datos. Esta idea proviene de una extensión intuitiva del teorema del límite central clásico. En particular, su uso no necesita estructuras de datos especiales, no necesita mantener distancias calculadas en la memoria y no requiere asignaciones exhaustivas repetidas. Está demostrado que el uso de K-means-lite obtiene una drástica ganancia de eficiencia y puede resolver grandes conjuntos de datos en tiempo real; En este artículo se denomina agrupación de datos avanzada (ADC) [22].

2.2. Representación de las incertidumbres de entrada

En este artículo, para tener en cuenta las incertidumbres relativas a la potencia de salida de generadores , cargas, etc., y los parámetros de los componentes, se representan mediante distribuciones probabilísticas. Con base en sus datos históricos, se pueden estimar las distribuciones. También pueden proporcionarse mediante una técnica de previsión, especialmente en la resolución de problemas operativos.

• Generación eólica Para

modelar la velocidad del viento, se utiliza comúnmente la distribución de Weibull [23] . Su función de densidad de probabilidad (PDF) se representa como:

$$f(v) = \frac{h}{c} \cdot \frac{v}{c}^{h-1} \cdot \exp\left(-\frac{v}{c}^{h}\right) \quad (1)$$

donde h: el parámetro de forma; c: el parámetro de escala; v: velocidad del viento.

La curva característica de la turbina eólica se puede estimar mediante la relación potencia eólica-velocidad del viento. pares de datos de medición [24]. También se puede modelar mediante una función por partes de la siguiente manera:

$$P_{wo}(v) = \begin{cases} 0 & v \le v_{ci} \text{ o } v > v_{co} \\ \dfrac{v - v_{ci}}{P_{wvr} - v_{ci}} & v_{ci} < v \le v_r \\ \text{encendido} & v_r < v \le v_{co} \end{cases} \quad (2)$$

donde vci, vco y vr son la velocidad del viento de activación, desactivación y velocidad nominal, respectivamente; Pwr y Pwo son la potencia nominal y la producción de la generación eólica, respectivamente.

• Generación solar

La distribución de la radiación solar se puede estimar a partir de los datos observados. También suele ser representado por una distribución Beta [25], como:

$$f(r) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \left(\frac{r}{rmax}\right)^{\alpha-1} \cdot \left(1 - \frac{r}{rmax}\right)^{\beta-1} \quad (3)$$

donde r y rmax son las radiaciones solares real y máxima, respectivamente; α y β son dos parámetros principales de la distribución; Γ(·) es la conocida función Gamma.

$$Pv(r) = \begin{cases} Pvr \dfrac{2r}{rcrstd\ r} & r < rc \\ pv \dfrac{}{primero} & rc \le r \le primero \\ PVR & r > primero \end{cases} \quad (4)$$

donde rc es la radiación en un punto determinado; rstd es la radiación estándar (correspondiente al entorno estándar); Pvr y Pvo son las potencias nominal y de salida de la unidad fotovoltaica, respectivamente. Comúnmente se requiere que la generación solar funcione en el modo de factor de potencia unitario, es decir, que su potencia reactiva sea igual a cero.

• Carga

La incertidumbre de cada carga suele estar representada por una distribución gaussiana o normal [1]. La función de distribución normal es una función continua y una de las funciones más utilizadas en la mayoría de los campos.

El PDF de una distribución normal es el siguiente:

$$f(x) = \sqrt{\frac{1}{2\pi\sigma}}\ mi^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (5)$$

en donde μ y σ son la expectativa (valor promedio) y la desviación estándar, respectivamente.

La función de distribución acumulada (CDF) de la función de distribución normal se calcula de la siguiente manera:

$$F(x) = \sqrt{\frac{1}{2\pi\sigma}} \int_{-\infty}^{x} mi^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt \quad (6)$$

Para modelar una carga, el valor esperado (media) es su potencia base, mientras que el estándar Se supone que la desviación es igual a un cierto porcentaje, por ejemplo, 10%, de la media.

Además de las populares distribuciones de probabilidad, mostradas arriba, que son muy adecuadas para representar incertidumbres de FER y cargas en sistemas eléctricos, actualmente, en los campos de probabilidad y estadística, existen otras distribuciones de probabilidad y funciones de densidad que se utilizan para incorporar incertidumbres en El problema de la FPP. En otras palabras, el hecho de que asumamos las funciones de distribución anteriores para FER y cargas no pierde la generalidad del uso del método PPF desarrollado en este estudio.

2.3. Flujo de energía probabilístico basado en agrupaciones de datos avanzados

El diagrama de flujo del enfoque propuesto, es decir, el flujo de potencia probabilístico basado en agrupamiento de datos avanzado (ADCPPF), utilizado para la evaluación de seguridad probabilística, se muestra en la Figura 1. En el diagrama de flujo de la Figura 1, la parte principal que ayuda a mejorar significativamente el tiempo de cálculo está en el bloque ADC.

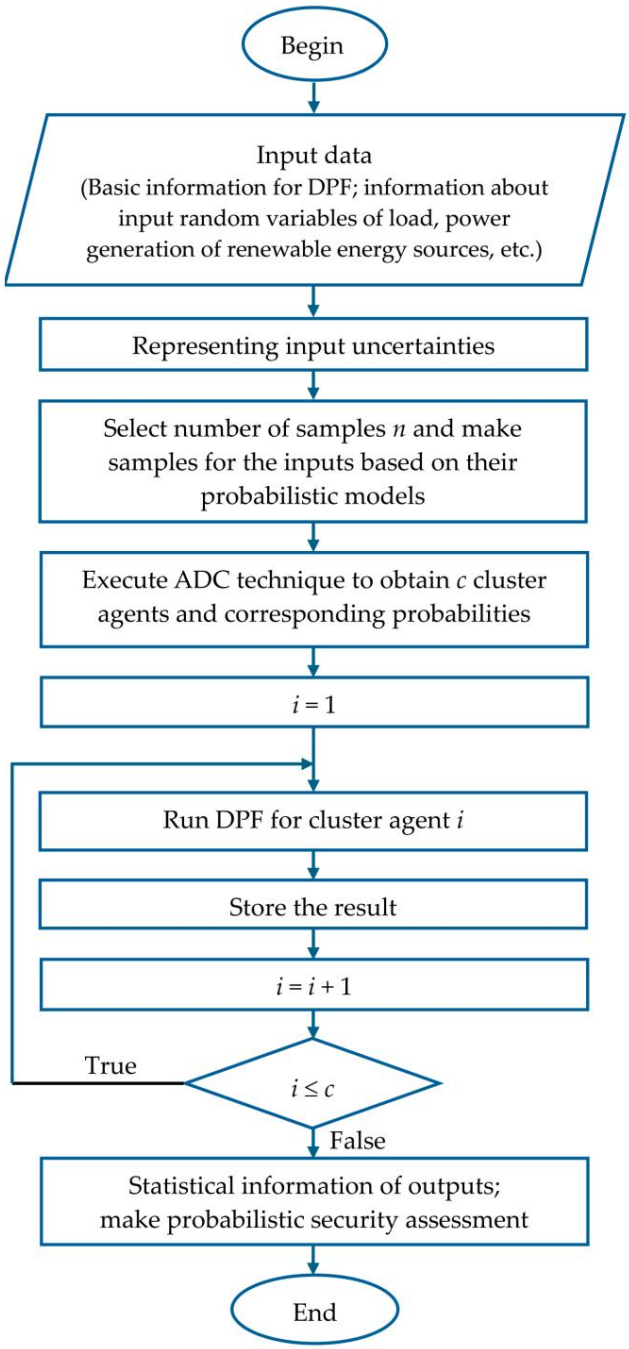


Figura 1. Diagrama de flujo de la evaluación de seguridad probabilística propuesta.

3. Pruebas y resultados

La aplicación del enfoque desarrollado se ilustra en un sistema de bus IEEE 300 modificado. La información necesaria para el análisis DPF del sistema, es decir, el diagrama de la red eléctrica, datos de barra, ramal y generador, se proporciona en [26]. El sistema incluye 300 autobuses, cargas y 69 generadores. En esta prueba, las incertidumbres de ambas cargas y de las RES. El sistema se modifica añadiendo 10 solares fotovoltaicas y 8 eólicas se muestra en las Tablas 1 y 2, respectivamente.

Tabla 1. Información de distribuciones Beta de energía solar fotovoltaica.

Autobús	Potencia nominal (MW) 50	Parámetro α	Parámetro 8
196		2.5	
198	80	1.6	9
203	40	1.2	6

Tabla 1. Información de distribuciones Beta de energía solar fotovoltaica.

Autobús	Potencia nominal (MW)	Parámetro α	Parámetro β
196	50	2.5	8
198	80	1.6	9
203	40	1.2	6
204	60	2.2	8
215	70	3.2	7
217	50	3.5	10
221	35	4.2	11
229	90	2.8	8
245	[illegible]	1.9	7
246	95	3.1	8

Tabla 2. Información sobre distribuciones Weibull de energía eólica.

Autobús	Potencia nominal (MW)	Parámetro de escala	Parámetro de forma
118	90	10	2.4
121	80	15	1.6
126	100	11	2.4
142	50	14	1.5
154	40	20	2.2
156	60	28	1.8
159	70	[illegible]	1.7
161	95	12	2.3

En aras de la simplicidad, pero sin pérdida de generalidad, las incertidumbres de las cargas y se supone que las FER son conocidas mediante técnicas de pronóstico. La incertidumbre de cada carga. se modela mediante una distribución normal con un valor esperado igual al valor base y desviación estándar igual al 10% del valor esperado. La energía solar fotovoltaica Se supone que la incertidumbre en cada planta sigue la distribución Beta, con sus parámetros dados. en la Tabla 1. También se supone que las distribuciones están correlacionadas con un coeficiente de correlación de 0,7. En este estudio, se supone que el escenario de simulación ocurre durante las horas del día, con condiciones potenciales para alguna generación de energía solar fotovoltaica. Además, nosotros asumir que las cargas y las plantas de energía solar y eólica no están bajo condiciones climáticas anormales condiciones (p. ej., ola de calor extrema, tormenta invernal a gran escala y huracanes), extremadamente fenómenos raros (p. ej., eclipse solar) y catástrofes (p. ej., guerras, terremotos y tsunamis). Algunos de estos eventos anormales son extremadamente difíciles de pronosticar con estándares bajos. desviación. Estos eventos anormales y raros también pueden describirse mediante una distribución adecuada. funciones e incluidas en el problema PPF. Sin embargo, esta cuestión está fuera del alcance de la estudio actual y se pretende que se considere en estudios futuros. De manera similar, la incertidumbre Se supone que la mayor parte de la producción de energía en cada parque eólico tiene distribuciones Weibull, con la parámetros que se muestran en la Tabla 2. Las distribuciones están correlacionadas con un coeficiente de correlación de 0,8.

En la prueba actual, se utiliza la potencia base de 100 MVA. Los resultados de MCS se utilizan como la referencia para evaluar los resultados obtenidos por otros métodos. Todas las pruebas se ejecutan en Matlab (R2015b) en una PC con CPU Intel Core i5 a 2,53 GHz y 4,00 GB de RAM.

El cálculo de PPF se realiza para lograr todos los resultados de interés de los RV de salida en términos de PDF y/o CDF. A modo de ilustración, las distribuciones de un número Se muestran los RV de salida seleccionados. La Figura 2 traza los CDF del PF real a través de la rama 2-8

Electrónica 2024, 13, x PARA REVISIÓN POR PARES 8 de 13

(es decir, denotado como P2–8), mientras que el voltaje en el bus 89 (es decir, denotado como V89) se da en la figura 3. De las figuras se puede ver que el enfoque ADCPPF desarrollado puede igualar bien con las curvas de MCS, lo que indica el buen desempeño del enfoque ADCPPF.

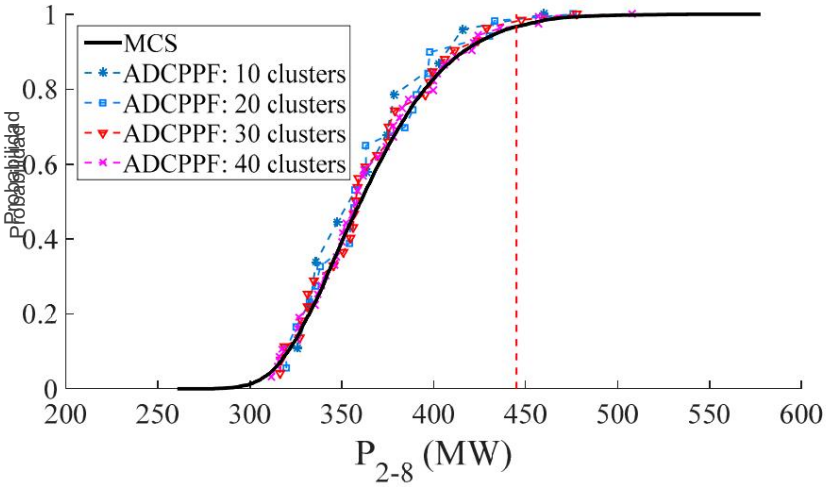


Figura 2. CDF de PF activo a través de la rama 2-8.
Figura 2. CDF de PF activo a través de la rama 2-8.
Figura 2. CDF de PF activo a través de la rama 2-8.

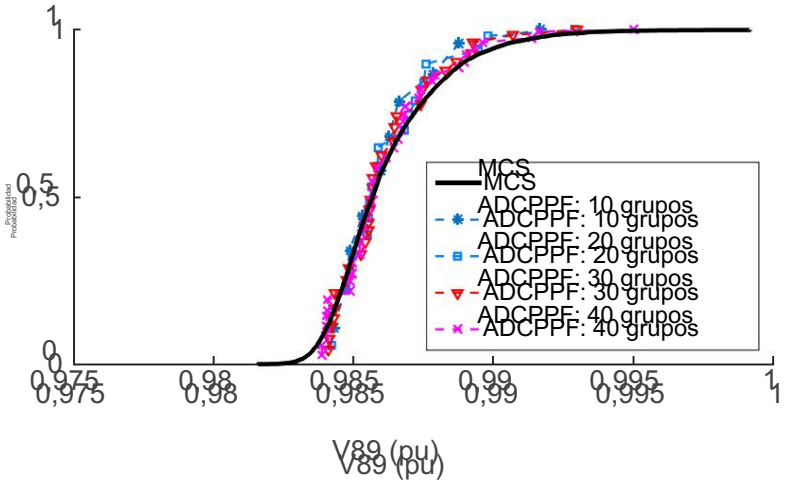


Figura 3. CDF de voltaje en el bus 89.
Figura 3. CDF de voltaje en el bus 89.
Figura 3. CDF de voltaje en el bus 89.

Porque es difícil observar con claridad al representar todos los resultados en la misma figura. Porque es difícil observar con claridad al representar todos los resultados en la misma figura, solo se representan los resultados correspondientes al método ADCPPF. Sin embargo, para figurar, sólo se representan los resultados correspondientes al método ADCPPF. Sin embargo, para demostrar su efectividad, también comparamos los resultados de PPF usando ADC con K- para demostrar su eficacia, también comparamos los resultados de PPF utilizando ADC con agrupación de K-medias.
Para demostrar su eficacia, también comparamos los resultados de PPF utilizando ADC con agrupación de K-medias.
Significa agrupamiento.
K-significa agrupación.

Como se analizó en la Sección 2.1, la agrupación de K-medias puede consumir mucho tiempo con grandes conjuntos de datos debido al problema de escasa escalabilidad. A diferencia de K-means, el enfoque K-means-lite se desarrolló en [16] y puede dar resultados muy rápidamente. Su precisión también aumenta en comparación con la técnica K-means. Por lo tanto, se ha demostrado que esto aumenta en precisión en comparación con el método K-medias. En cambio, se centra en el rendimiento del tiempo de procesamiento, lo que indica que el enfoque propuesto tiene buena aplicabilidad para resolver problemas en el marco temporal de operación del sistema eléctrico. En la Tabla 3 se muestra claramente que el enfoque ADCPPF puede impulsar para dar el resultado en unos pocos segundos, en comparación con los cientos de segundos que necesita MCS. En particular, como se mencionó anteriormente, para un conjunto de datos que incluye n puntos de datos, K-MCS. En particular, como se mencionó anteriormente, para un conjunto de datos que incluye n puntos de datos, la agrupación K-medias necesita una cantidad de tiempo igual a O(nkT), donde T es el número de iteraciones). El tiempo de ejecución de K-medias aumentará drásticamente con un aumento de n y/o k. El sistema de bus IEEE 300 modificado es un sistema a gran escala, por lo que el PPF basado en K-means funciona muy duro y lleva mucho tiempo.

Como se analizó en la Sección 2.1, la agrupación de K-medias puede consumir mucho tiempo con grandes conjuntos de datos debido al problema de escasa escalabilidad. A diferencia de K-means, el enfoque K-means-lite se desarrolló en [16] y puede dar resultados muy rápidamente. Su precisión también aumenta en comparación con la técnica K-means. Por lo tanto, se ha demostrado que esto aumenta en precisión en comparación con el método K-medias. En cambio, se centra en el rendimiento del tiempo de procesamiento, lo que indica que el enfoque propuesto tiene buena aplicabilidad para resolver problemas en el marco temporal de operación del sistema eléctrico. En la Tabla 3 se muestra claramente que el enfoque ADCPPF puede impulsar para dar el resultado en unos pocos segundos, en comparación con los cientos de segundos que necesita MCS. En particular, como se mencionó anteriormente, para un conjunto de datos que incluye n puntos de datos, K-MCS. En particular, como se mencionó anteriormente, para un conjunto de datos que incluye n puntos de datos, la agrupación K-medias necesita una cantidad de tiempo igual a O(nkT), donde T es el número de iteraciones). El tiempo de ejecución de K-medias aumentará drásticamente con un aumento de n y/o k. El sistema de bus IEEE 300 modificado es un sistema a gran escala, por lo que el PPF basado en K-means funciona muy duro y lleva mucho tiempo.

k. El sistema de bus IEEE 300 modificado es un sistema a gran escala, por lo que se ejecuta PPF basado en K-means. muy duro y lleva mucho tiempo.

Tabla 3. Comparación de tiempos de ejecución.

Método	Tiempo (s)
MCS	725
ADCPPF con 5 grupos	2.64
ADCPPF con 10 grupos	2,79
ADCPPF con 20 grupos	2,98
ADCPPF con 30 grupos	3.26
ADCPPF con 40 grupos	3.47
ADCPPF con 50 clusters	3.72
ADCPPF con 70 clusters	4.21
PPF basada en K-medias con 5 grupos	140
PPF basado en K-medias con 10 grupos	420

A partir de las Figuras 2 y 3 y la Tabla 3, la precisión de ADCPPF aumenta (es decir, intuitivamente, la curva correspondiente en las Figuras 2 y 3 sigue más de cerca la curva MCS) y El tiempo requerido para ejecutar el método también aumenta con un número creciente de racimos. Al comparar MCS usando K-means y ADCPPF, K-means es un desafío en este caso y hace que el tiempo de cálculo aumente mucho más que cuando se utiliza ADCPPF. A través del análisis anterior, se muestra que el método ADCPPF tiene la ventaja de tanto una precisión relativamente alta como un tiempo de ejecución significativamente reducido.

PPF puede proporcionar distribuciones para vehículos recreativos de salida que sean buenas para la seguridad del sistema de energía. evaluación. La probabilidad de subtensión o sobretensión, sobrecarga de línea, etc., puede ser juzgado. Por ejemplo, se supone que el límite superior del PF real de la rama 2 a 8 es igual a 445 MW, es decir, correspondiente a la línea vertical de la Figura 2, y la probabilidad de que la potencia transmitido a través de la rama supera su límite se puede calcular como:

$$P\{P2-8 > 445\} = 1,4\% \quad (7)$$

De manera similar, la probabilidad de que el voltaje en un bus considerado esté fuera del rango operativo puede ser evaluado. Sin embargo, en esta prueba, los voltajes en todos los buses del sistema (por ejemplo, V89 en la Figura 3) están dentro del rango, es decir, [0,9, 1,1] pu

Los resultados obtenidos por el método ADCPPF pueden ayudar al operador del sistema a evaluar los estados de operación para tomar decisiones adecuadas y brindar soluciones el sistema.

4. Discusión adicional sobre la aplicabilidad de varios métodos PPF

Los métodos PPF se pueden seleccionar para su aplicación tanto en problemas de planificación como de operación. de los sistemas de energía.

- Aplicación de PPF en problemas de planificación: el método MCS es adecuado para resolver problemas de planificación. problemas con períodos de tiempo prolongados (como años, estaciones, meses, semanas) o operativos. problemas de planificación en el plazo de unos pocos días. En tales casos, el tiempo para lograr Los resultados no tienen por qué ser muy rápidos. Además de los datos de configuración de la red, los datos sobre fuentes (especialmente FER) y cargas recolectadas durante largos períodos, es decir, meses, un Se utilizan un año o varios años para estimar las PDF. Estos datos también pueden ser utilizados por un técnica de pronóstico para proporcionar resultados para la operación del sistema. Si el pronóstico técnica sigue el enfoque de pronóstico puntual, los resultados del pronóstico se proporcionan como un valor establecido en cada momento del pronóstico y un error correspondiente. Estos valores son consideró la expectativa y la desviación estándar de la función de distribución normal.

Electrónica 2024, 13, x PARA REVISIÓN POR PARES 10 de 13

Estas funciones son la información de entrada del problema MCS. Si el pronóstico
La técnica sigue un enfoque de pronóstico de incertidumbre como el pronóstico probabilístico.
o enfoques de pronóstico de escenarios [24], entonces la función de distribución de probabilidad será más útil. De hecho, cuando se aplica MCS en la práctica, si el sistema es demasiado grande, con muchas variables de entrada, el procesamiento de MCS es muy difícil, ocupando un tiempo extremadamente grande cantidad de memoria, lo que hace que sea muy difícil de procesar y, en muchos casos, puede que no sea posible. En tales casos, la agrupación (como en el artículo actual) y la dimensionalidad se deben utilizar técnicas de reducción de dimensionalidad.

• Aplicación en problemas operacionales: Para problemas operacionales con tiempos extremadamente cortos marcos de tiempo que van desde unos pocos minutos (es decir, un marco de muy corto plazo) hasta unas pocas horas (es decir, marco de corto plazo) y dentro de 24 h (día por delante), métodos analíticos y aproximación Se pueden aplicar métodos de información con un tiempo de procesamiento rápido. El método MCS también se puede utilizar en un marco de tiempo de varias horas o más cuando el sistema de energía es de pequeña escala. Combinado y el número de RV de entrada también es pequeño. El método MCS combinado con técnicas de agrupamiento, como la propuesta en este artículo, se pueden implementar con técnicas de previsión con plazos de funcionamiento extremadamente cortos. Basado en datos de RES (estaciones de RES o cargas (estaciones de red), distribución de probabilidad), cargas probables Se construyen funciones de distribución de probabilidad para los puntos temporales del problema operativo. Se Estas funciones proporcionan entradas por una técnica de pronóstico.

La siguiente parte es un ejemplo ilustrativo. Supongamos una técnica para pronosticar la carga. La siguiente parte de energía de las FER proporciona información de pronóstico en diferentes momentos del plazo operativo. En ese momento, los resultados obtenidos con los métodos PPF serán muy útil para el operador del sistema.

La Figura 4 ilustra la PDF de la corriente que fluye por una línea de interés en periodos consecutivos. veces (resolución de 1 h) en el horizonte temporal de 24 h de funcionamiento diario del sistema. El operador del sistema tendrá una "imagen" muy clara del riesgo de seguridad del sistema, por lo que pueden proponer soluciones adecuadas. Sobre la base de comparar la curva PDF con la con el límite máximo Imax (correspondiente al límite de potencia) de la línea de interés, es posible determinar el punto en el que existe el riesgo de que la corriente que pasa por esa línea supere el valor permitido.

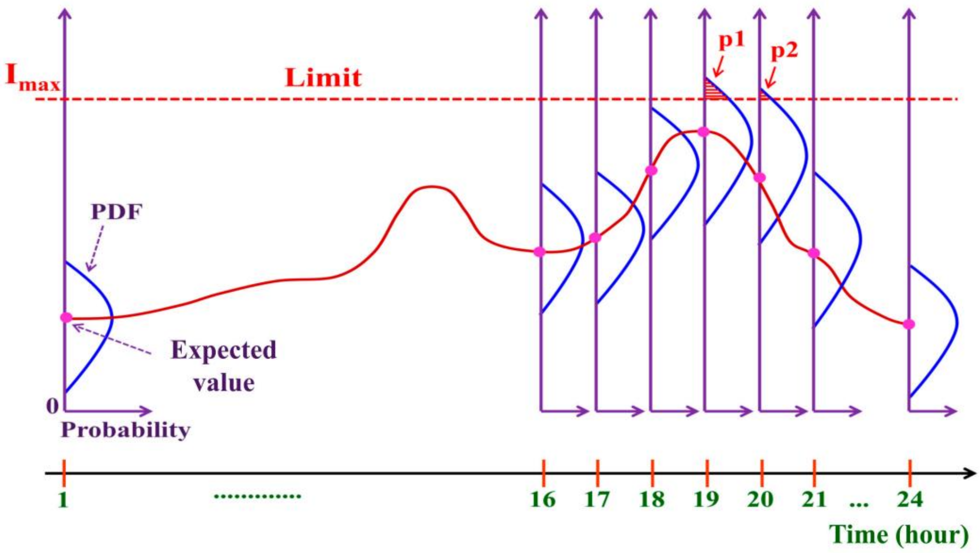


Figura 4. El ejemplo ilustra la PDF de corriente que fluye en una línea interesada en consecutivas veces en el horizonte temporal de 24 h de funcionamiento diario del sistema.

Por ejemplo, en la Figura 4, de 16:00 a 18:00, la corriente aumenta gradualmente y, cerca de las 19:00, el límite superior de la línea PDF comienza a tocar y superar el límite Imax. y la probabilidad correspondiente a las 19:00 horas es p1. Cabe señalar que, como la probabilidad Al anochecer, la fuente de energía solar disminuye gradualmente y, en este ejemplo, a las 19:00, esta fuente ya no genera energía. Sin embargo, este es el período pico en el sistema donde la carga aumenta rápidamente hasta alcanzar el pico. En este momento, la incertidumbre en el sistema proviene de las cargas y otras fuentes, si las hay, y no de la fuente de energía solar. Luego, a las 20:00 horas, la corriente tiende a disminuir y el nivel de intrusión ha aumentado.

No se ven más pequeños ni em máx ni saey rss ue oe overoang. Es un punto muy nuevo desde el punto de vista de la evaluación de la seguridad del sistema.

La Figura 5 es un ejemplo ilustrativo de la evaluación del riesgo de sobre/subtensión en un autobús de interés. En este ejemplo, el riesgo de subtensión a las 19:00 y 20:00 se calcula como p1 y p2, respectivamente.

El método de cálculo y análisis de la seguridad del sistema se basa en información y datos obtenidos de cantidades de entrada aleatorias del problema, como las cargas y potencias de salida de las RES. El resultado es que, en términos de distribuciones de probabilidad de la producción al aproximarse por la noche, la fuente de energía solar disminuye gradualmente y, en este ejemplo, en cantidades como la tensión del nodo o la potencia o corriente transmitida en las ramas. Esta 19:00 horas, esta fuente ya no genera energía. Sin embargo, este es el período pico en el sistema. El proceso se realiza antes de la operación real para encontrar las distribuciones de probabilidad de donde la carga aumenta rápidamente hasta alcanzar el pico. En este momento, la incertidumbre en cantidades de interés. el sistema proviene de las cargas y otras fuentes, si las hubiera, y no de la energía solar.

Luego, a las 20:00 horas, la corriente sigue en disminución. Para los sistemas de energía reales con un sistema SCADA EMS, este sistema proporcionará fuente de información. formación sobre los parámetros del modo y se actualizará periódicamente casi en tiempo real, por lo que mon- de p2. Así, en la supervisión del funcionamiento de los parámetros de modo grados se realiza de forma continua. Además dura. Además, cuando existe un sistema SCADA EMS, los datos recopilados para factores aleatorios serán más convenientes y se actualizarán continuamente. Cuando el conjunto de datos de factores aleatorios es más una técnica de pronóstico que completo, la información obtenida es más clara. Ese también es otro beneficio cuando se utiliza una sobrecarga considerada Sistema SCADA EMS combinado con el método PPF. cuando existe riesgo de sobrecarga durante un largo período de tiempo y el nivel de sobrecarga es

El objetivo principal de este documento es desarrollar un enfoque para calcular el PPF que garantice que el operador una cierta precisión aceptable pero que además proporciona resultados muy rápidos para cumplir con lo deseado. Cabe señalar que, en la Figura 4, si solo nos preocupamos por el valor esperado (similar a aplicación en el marco temporal casi "real" de operación del sistema eléctrico. Como los resultados obtenidos del DPF tradicional), todos los valores actuales calculados son más pequeños. Como se mencionó anteriormente, el método MCS enfrenta muchos desafíos e incluso es imposible cuando el límite Imax y los riesgos de seguridad debidos a la sobrecarga "no se ven". Esto es muy aplicado a sistemas de energía reales, especialmente sistemas a gran escala con un funcionamiento muy corto, un nuevo punto desde el punto de vista de la evaluación de la seguridad del sistema. plazos. El algoritmo de agrupamiento propuesto para resolver el problema de PPF. La Figura 5 es un ejemplo ilustrativo de la evaluación del riesgo de sobre/subtensión en un autobús en este documento es simple, fácil de implementar y ayuda a abordar eficazmente la mala escala de interés. En este ejemplo se calcula el riesgo de subtensión en los horarios de las 19:00 y las 20:00 habilidad del algoritmo K-medias. Por lo tanto, ADCPPF puede dar resultados rápidos con datos grandes como p1 y p2, respectivamente. que se puede aplicar para resolver problemas de PPF para sistemas de energía a gran escala.

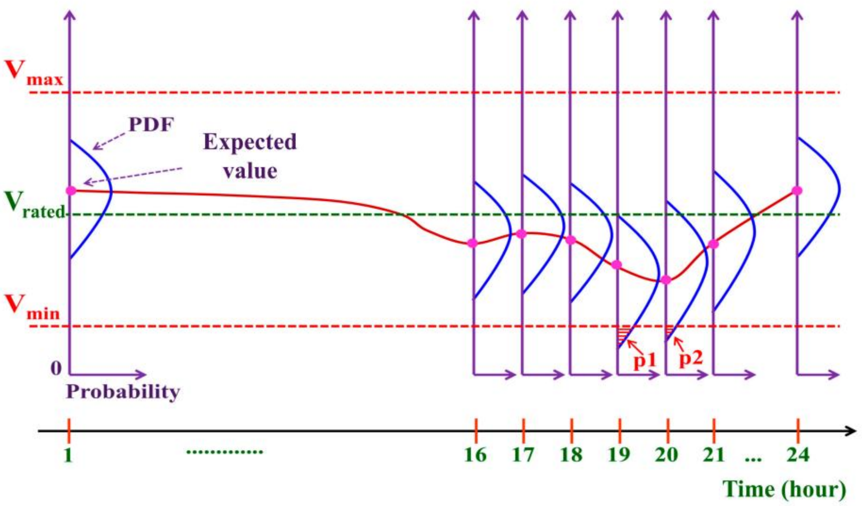


Figura 5. El ejemplo ilustra la PDF del voltaje en un nodo de interés en momentos consecutivos en el horizonte temporal de 24 h de operación diaria del sistema eléctrico.

El método de cálculo y análisis de la seguridad del sistema se basa en información y datos obtenidos de cantidades de entrada aleatorias del problema, como la cargas y potencias de salida de las FER. El resultado es en términos de distribuciones de probabilidad de cantidades de salida como el voltaje del nodo o la potencia o corriente transmitida en las ramas. Este proceso se realiza antes de la operación real para encontrar las distribuciones de probabilidad de las cantidades de interés.

Para sistemas de energía reales con un sistema SCADA EMS, este sistema proporcionará información sobre los parámetros del modo y se actualizará periódicamente casi en tiempo real, para que El seguimiento del funcionamiento de los parámetros de modo grados se realiza de forma continua. En Además, cuando exista un sistema SCADA EMS, los datos recopilados para factores aleatorios serán más conveniente y continuamente actualizado. Cuando el conjunto de datos de factores aleatorios es más completo, la información obtenida es más clara. Ese es también otro beneficio al usar un Sistema SCADA EMS combinado con el método PPF.

El objetivo principal de este documento es desarrollar un enfoque para calcular la FPP que garantice cierta precisión aceptable pero que también proporciona resultados muy rápidos para cumplir con los objetivos deseados.

aplicación en el marco temporal casi “real” de operación del sistema eléctrico. Como se mencionó anteriormente, el método MCS enfrenta muchos desafíos e incluso es imposible cuando se aplica a sistemas de energía reales, especialmente sistemas a gran escala con tiempos de operación muy cortos . El algoritmo de agrupamiento propuesto para resolver el problema PPF en este artículo es simple, fácil de implementar y ayuda a abordar de manera efectiva la escasa escalabilidad del algoritmo K-means. Por lo tanto, ADCPPF puede brindar resultados rápidos con grandes datos que pueden aplicarse para resolver problemas de PPF para sistemas de energía a gran escala.

Sin embargo, además de las ventajas y contribuciones del método propuesto en aplicaciones prácticas, también tiene una limitación que debe abordarse: predeterminar el valor de k como en el algoritmo K-means. Además, las limitaciones inherentes del PPF, que no cubre los transitorios electromagnéticos, y las condiciones operativas anormales extremas son cuestiones sin resolver. La planificación y operación del sistema eléctrico también debe centrarse en los peores escenarios futuros y anormales (por ejemplo, olas de calor, tormentas invernales, etc.). Los fenómenos meteorológicos extremos exacerbados por el cambio climático y el calentamiento global plantean un alto nivel de incertidumbre. Por lo tanto, se debe considerar una mayor diversidad en los tipos de incertidumbre para la integración en el problema de la FPP. Los sistemas de almacenamiento de energía (por ejemplo, baterías, energía hidroeléctrica de almacenamiento por bombeo y esquemas de conexión de vehículos eléctricos a la red) pueden mitigar la incertidumbre relacionada con las FER que tampoco se considera en este estudio. El hardware es uno de los factores importantes que afectan el análisis PPF, especialmente en relación con el tiempo de cálculo. Las investigaciones futuras también pueden centrarse en el procesamiento del algoritmo en computadoras de alto rendimiento, superando los problemas de procesamiento del tiempo y permitiendo centrarse más en la precisión del modelo. Estas limitaciones abren temas a considerar en futuras invest

5. Conclusiones

PPF es una herramienta eficaz para calcular y analizar sistemas de energía y considerar las incertidumbres existentes en el sistema. Puede ayudar al operador del sistema a evaluar la seguridad. Entre las diversas técnicas para PPF, MCS proporciona resultados muy precisos pero a menudo requiere un uso intensivo de cálculo. Este artículo se centra en resolver el problema de reducir el tiempo de cálculo para la simulación Monte Carlo para lograr una herramienta práctica con alta precisión que proporcione resultados de cálculo rápidos que se utilizarán para resolver problemas en el horizonte temporal de las operaciones del sistema eléctrico. Para lograr ese objetivo, utilizamos una técnica avanzada de agrupación de datos llamada K-means-lite. El enfoque desarrollado, ADCPPF, se prueba exhaustivamente en un sistema de bus IEEE-300 modificado y muestra un buen rendimiento en la reducción del tiempo de cálculo.

Contribuciones de los autores: Metodología, NTAN, DDL, VDN, VKP y VKH; software, NTAN y DDL; Todos los autores escribieron y editaron el manuscrito. Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: Esta investigación fue financiada por el Ministerio de Educación y Capacitación de Vietnam con el número de proyecto CT2022.07.DNA.03.

Declaración de disponibilidad de datos: los datos están contenidos en el artículo.

Conflictos de intereses: Los autores declaran no tener conflictos de intereses.

Referencias

1. Le, DD; Berizzi, A.; Bovo, C. Un enfoque de evaluación probabilística de la seguridad de los sistemas de energía con recursos eólicos integrados. Renovar. Energía 2016, 85, 114-123. [Referencia cruzada]

2. Borkowska, B. Flujo de carga probabilístico. Traducción IEEE. Aplicacion de energia Sistema. 1974, PAS-93, 752–759. [Referencia cruzada]

3. Allan, enfermero registrado; Al-Shakarchi, MRG Técnicas probabilísticas en análisis de flujo de carga de CA. Proc. Inst. Electo. Ing. 1977, 124, 154-160. [Referencia cruzada]

4. Zhang, P.; Lee, ST Cálculo probabilístico del flujo de carga utilizando el método de acumuladores combinados y expansión de Gram-Charlier. Traducción IEEE. Sistema de energía. 2004, 19, 676–682. [Referencia cruzada]

5. Fan, M.; Vittal, V.; Heydt, GT; Ayyanar, R. Estudios probabilísticos de flujo de energía para sistemas de transmisión con energía fotovoltaica Generación mediante acumuladores. Traducción IEEE. Sistema de energía. 2012, 27, 2251–2261. [Referencia cruzada]

6. Le, DD; Berizzi, A.; Bovo, C.; Ciapessoni, E.; Cirio, D.; Pitto, A.; Gross, G. Un enfoque probabilístico para la evaluación de la seguridad del sistema eléctrico en condiciones de incertidumbre. En Actas de Bulk Power System Dynamics and Control—IX Optimization, Security and Control of the Emerging Power Grid (Simposio IREP de 2013), Creta, Grecia, 25 a 30 de agosto de 2013.
7. Su, CL Cálculo probabilístico del flujo de carga utilizando el método de estimación puntual. Traducción IEEE. Sistema de energía. 2005, 20, 1843–1851. [Referencia cruzada]
8. Mohammadi, M.; Shayegani, A.; Adaminejad, H. Un nuevo enfoque del método de estimación puntual para el flujo de carga probabilístico. En t. J. eléctrico. Poder 2013, 51, 54–60. [Referencia cruzada]
9. Liu, Y.; Gao, S.; Cui, H.; Yu, L. Flujo de carga probabilístico considerando correlaciones de variables de entrada siguiendo distribuciones arbitrarias. eléctrico. Sistema de energía. Res. 2016, 140, 354–362. [Referencia cruzada]
10. Yu, H.; Chung, CY; Wong, KP; Lee, HW; Zhang, JH Evaluación probabilística del flujo de carga con muestreo de hipercubo latino híbrido y descomposición de Cholesky. Traducción IEEE. Sistema de energía. 2009, 24, 661–667. [Referencia cruzada]
11. Zou, B.; Xiao, Q. Resolución del problema probabilístico del flujo de energía óptimo utilizando el método Cuasi Monte Carlo y el noveno orden Transformación normal polinómica. Traducción IEEE. Sistema de energía. 2014, 29, 300–306. [Referencia cruzada]
12. Hajian, M.; Rosehart, WD; Zareipour, H. Flujo de energía probabilístico mediante simulación de Monte Carlo con muestreo de Supercubo Latino. Traducción IEEE. Sistema de energía. 2013, 28, 1550–1559. [Referencia cruzada]
13. Leite da Silva, AM; Milhorance de Castro, A. Evaluación de riesgos en flujo de carga probabilístico mediante simulación de Monte Carlo y Método de entropía cruzada. Traducción IEEE. Sistema de energía. 2018, 14, 1193–1202. [Referencia cruzada]
14. Carpinelli, G.; Caramia, P.; Varilone, P. Método de simulación multilineal de Monte Carlo para flujo de distribución de carga probabilística sistemas con sistemas de generación eólica y fotovoltaica. Renovar. Energía 2015, 76, 283–295. [Referencia cruzada]
15. Hagh, MT; Amiyán, P.; Galvani, S.; Valizadeh, N. Flujo de carga probabilístico utilizando la agrupación de optimización de enjambre de partículas método. IET Gen. Trans. Distribuir. 2018, 12, 780–789. [Referencia cruzada]
16. Khalghani, señor; Ramezani, M.; Mashhadi, MR Flujo de energía probabilístico basado en simulación Monte-Carlo y agrupación de datos para analizar sistemas de energía a gran escala incluyendo parques eólicos. En Actas de la Conferencia IEEE Kansas Power and Energy (KPEC) de 2020, Manhattan, KS, EE. UU., 13 y 14 de julio de 2020.
17. Hachís, MS; Hasanien, HM; Ji, H.; Alkuhayli, A.; Alharbí, M.; Akmaral, T.; Turquía, RA; Jurado, F.; Badr, AO Monte Carlo Simulación y técnica de agrupación para resolver el problema probabilístico del flujo de energía óptimo para sistemas híbridos de energía renovable. Sostenibilidad 2023, 15, 783. [CrossRef]
18. Zhang, R.; Chen, Z.; Li, Z.; Jiang, T.; Li, X. Funcionamiento robusto en dos etapas de microrredes multienergéticas integradas de electricidad, gas y calor. considerando incertidumbres heterogéneas. Aplica. Energía 2024, 371, 123690. [CrossRef]
19. Jiang, Y.; Ren, Z.; Li, W. Región de operación de emisiones de carbono comprometida para sistemas energéticos integrados: conceptos y análisis. Traducción IEEE. Sostener. Energía 2024, 15, 1194–1209. [Referencia cruzada]
20. Li, Z.; Xu, Z.; Wang, P.; Xiao, G. Restauración de un sistema de distribución de energía múltiple con reconfiguración de la red de distrito conjunta mediante programación estocástica distribuida. Traducción IEEE. Red inteligente 2024, 15, 2667–2680. [Referencia cruzada]
21. Miraftabzadeh, SM; Colombo, CG; Longo, M.; Foiadelli, F. K-Means y métodos de agrupación alternativos en sistemas de energía modernos. Acceso IEEE 2023, 11, 119596–119633. [Referencia cruzada]
22. Olukanmi, PO; Nelwamondo, F.; Marwala, T. k-Means-Lite: Agrupación en tiempo real para grandes conjuntos de datos. En Actas de la Quinta Conferencia Internacional sobre Computación Suave e Inteligencia Artificial, Nairobi, Kenia, 21 y 22 de noviembre de 2018; págs. 54–59.
23. Wang, X.; Chiang, HD; Wang, J.; Liu, H.; Wang, T. Análisis de estabilidad a largo plazo de sistemas eléctricos con energía eólica basado en ecuaciones diferenciales estocásticas: desarrollo de modelos y fundamentos. Traducción IEEE. Sostener. Energía 2015, 6, 1534–1542. [Referencia cruzada]
24. Le, DD; Bruto, G.; Berizzi, A. Modelado probabilístico de la producción de parques eólicos en múltiples sitios para aplicaciones basadas en escenarios. IEEE Trans. Sostener. Energía 2015, 6, 748–758. [Referencia cruzada]
25. Salameh, ZM; Borowy, BS; Amin, ARA Emparejamiento módulo fotovoltaico-sitio basado en factores de capacidad. Traducción IEEE. Conversaciones de energía . 1995, 10, 326–332. [Referencia cruzada]
26. Archivo de casos de prueba del sistema de energía. Disponible en línea: http://labs.ece.uw.edu/pstca/pf300/pg_tca300bus.htm (consultado el 6 junio de 2024).

electrónica

MDPI

Artículo

Un esquema rápido de imágenes de radar de apertura sintética inversa Combinando disparos acelerados por GPU y rebote Ray y Algoritmo de retroproyección en anchos de banda y ángulos amplios

Jiongming Chen [1], Pengju Yang [1,2,*], Rong Zhang [1] y Rui Wu [1,3]

[1] Escuela de Física e Información Electrónica, Universidad de Yan'an, Yan'an 716000, China; jmchen@yau.edu.cn (JC); zr2108700733@yau.edu.cn (RZ); wurui@yau.edu.cn (RW)
[2] Laboratorio clave para la ciencia de la información de ondas electromagnéticas (MoE), Universidad de Fudan, Shanghai 200433,
[3] Instituto de Tecnología y Software de Radiofrecuencia de China, Instituto de Tecnología de Beijing, Beijing 100081, China
* Correspondencia: pjyang@fudan.edu.cn

Resumen: Las técnicas de imágenes de radar de apertura sintética inversa (ISAR) se utilizan con frecuencia en aplicaciones de clasificación y reconocimiento de objetivos, debido a su capacidad para producir imágenes de alta resolución para objetivos en movimiento. Para satisfacer la demanda de imágenes ISAR para cálculos electromagnéticos con alta eficiencia y precisión, se presenta un novedoso método de disparo y rebote de rayos acelerados (SBR) que combina una unidad de procesamiento de gráficos (GPU) y una estructura de árbol de jerarquías de volumen delimitador (BVH) . Para superar el problema de las imágenes desenfocadas mediante un procedimiento ISAR basado en Fourier en condiciones de gran angular y ancho de banda amplio, se desarrolla un algoritmo de imágenes de retroproyección paralela (BP) eficiente mediante la utilización de la técnica de aceleración de GPU. El SBR acelerado por GPU presentado se valida en comparación con el método RL-GO en el software comercial FEKO v2020. Para las imágenes ISAR, se indica claramente que se pueden observar fuertes centros de dispersión, así como perfiles de objetivos, bajo grandes ángulos de , y amplios anchos de banda, 3 GHz. Es acimut de observación; Δφ= 90◦ también indicó que las imágenes ISAR son muy sensibles a los ángulos de observación. Además, se pueden observar lóbulos laterales obvios, debido a que la historia de fase de la onda electromagnética se distorsiona como resultado de la dispersión multipolar. Los resultados de la simulación confirman la viabilidad y eficiencia de nuestro esquema al combinar SBR acelerado por GPU con el algoritmo BP para una simulación rápida de imágenes ISAR en condiciones de gran angular y ancho de banda amplio.

Palabras clave: imágenes ISAR; método de disparo y rebote de rayos; aceleración de GPU; árbol BVH; algoritmo de retroproyección

Cita: Chen, J.; Yang, P.; Zhang, R.; Wu, R. Un esquema rápido de imágenes de radar de apertura sintética inversa que combina un algoritmo de retroproyección y disparo acelerado por GPU y rayos de rebote en anchos de banda y ángulos amplios. Electrónica 2024, 13, 3062. https://doi.org/

10.3390/electrónica13153062

Editor académico: Chiman Kwan

Recibido: 4 de junio de 2024
Revisado: 29 de julio de 2024
Aceptado: 30 de julio de 2024
Publicado: 2 de agosto de 2024

1. Introducción

El radar de apertura sintética inversa (ISAR) es un potente sistema de radar activo de imágenes de microondas ampliamente utilizado en aplicaciones militares y civiles debido a su capacidad de producir imágenes de alta resolución para objetivos en movimiento en casi cualquier condición climática y durante todo el día [1–3] . Las imágenes ISAR se pueden obtener enfocando los datos del campo de dispersión en múltiples ángulos y frecuencias, que son una representación bidimensional del centro de dispersión del objetivo [4–6]. La simulación de imágenes ISAR para objetivos eléctricos de gran tamaño requiere mucho tiempo debido al cálculo de múltiples ángulos y campos de dispersión de frecuencias. Se han desarrollado varios métodos y sus versiones mejoradas, así como técnicas de aceleración, para calcular eficientemente la dispersión de objetivos eléctricos grandes, incluidos tanto el método numérico de baja frecuencia como los métodos de aproximación de alta frecuencia [7,8]. Debido a los enormes requisitos de memoria y tiempo de cálculo, los métodos numéricos puros, como el método de momentos (MoM) y el método de elementos finitos (FEM), se enfrentan a enormes desafíos [9]. Debido al buen compromiso entre precisión y eficiencia, los métodos de aproximación de alta frecuencia se utilizan ampliamente en la simulación de imágenes ISAR para

objetivos eléctricos grandes. Entre ellos, el método de rayos de disparo y rebote (SBR) es el más popular, que es una combinación de óptica física (PO) y óptica geométrica (GO) y es adecuado para tener en cuenta la dispersión múltiple.

Siguiendo la propuesta del método SBR por Ling en 1989, los investigadores han implementado numerosas mejoras [10], como el método de disparo y rebote de rayos en el dominio del tiempo (TDSBR), el trazado de rayos analítico bidireccional [11], etc. contribuyó a la creciente popularidad del método del rayo rebote. En los últimos años, se ha propuesto un método de rayos de rebote acelerado por GPU basado en un algoritmo transversal de árbol de dimensión k (Kd) sin pila , que permite que el proceso de trazado de rayos se lleve a cabo de manera eficiente en la GPU [12]. En [13], se propone un método mejorado de rayos que rebotan utilizando una técnica de propulsión de rayos para acelerar el proceso de trazado de rayos y mejorar la eficiencia de la intersección de rayos, permitiéndole calcular eficientemente las características de dispersión de objetivos eléctricamente grandes . En [14], se propone la inclusión de trayectorias de rayos inversos en el método SBR para mejorar la precisión de las predicciones de la sección transversal del radar de cavidad (RCS), que se pueden implementar en el código SBR existente de manera casi trivial, al tiempo que se producen mejoras potencialmente sustanciales en la precisión de la predicción. . Se propone una técnica de trazado de rayos inverso basada en la estructura de datos del árbol Kd de las cuerdas, que se ha demostrado que produce resultados satisfactorios en el cálculo de las características de dispersión de alta frecuencia [15]. Además de la mejora de SBR mediante la utilización de GPU y estructuras de datos, el método SBR también se ha integrado con otros métodos computacionales electromagnéticos, lo que hace que este método sea más completo en su consideración de las características de dispersión electromagnética de estructuras objetivo complejas [16- 19]. Por ejemplo, el método SBR basado en octree en combinación con la teoría física de la difracción (PTD) se presenta para el análisis de la dispersión electromagnética (EM) del objetivo en movimiento [20]. En [21], se propone un método híbrido de momento dipolar equivalente (EDM), MOM y SBR para mejorar la eficiencia computacional del RCS de objetos complejos dentro del marco EDM. En este método híbrido, se introduce un enfoque iterativo para mejorar el rendimiento del algoritmo, ofreciendo alta precisión y reduciendo el tiempo

Sobre la base del modelado de dispersión electromagnética, se pueden obtener imágenes ISAR enfocadas aplicando un algoritmo de procesamiento de señales, incluido el algoritmo Doppler de rango (RD), el algoritmo de formato polar (PFA), el algoritmo de retroproyección (BP), etc. [22-24]. El algoritmo de imágenes ISAR más utilizado interpola los datos polares en una cuadrícula cartesiana y luego aplica una FFT 2-D para lograr la reconstrucción ISAR. Como caso especial, en condiciones de ángulo pequeño y ancho de banda pequeño, las imágenes ISAR se pueden obtener aproximadamente realizando la transformada inversa de Fourier de datos de campo retrodispersados 2D, y las imágenes ISAR resultantes se componen de los centros de dispersión del objetivo con su coeficiente de reflexión electromagnética. . Debido a su idoneidad para el procesamiento paralelo de GPU y su capacidad para obtener imágenes ISAR en cualquier modo, el algoritmo BP y sus versiones modificadas se utilizan ampliamente en aplicaciones de imágenes SAR/ISAR [25]. Ya en 1989, se propuso un algoritmo BP simplificado y su arquitectura de procesamiento paralelo utilizando la forma de onda del radar como respuesta al impulso del filtro para obtener la proyección filtrada [26]. Hasta ahora, el algoritmo BP basado en GPU todavía se está desarrollando para optimizar el rendimiento máximo del algoritmo BP en servidores y dispositivos GPU miniaturizados, que pueden abordar las diferencias en las plataformas de hardware, así como las diferencias en las escalas de datos [27]. . En [28], se desarrolla un algoritmo de imágenes ISAR para escenas compuestas de objetivo-océano basado en disparos en el dominio del tiempo y rayos que rebotan (TDSBR). En [29], se propone la óptica física iterativa en el dominio del tiempo (TD-LIPO) para analizar la dispersión de objetivos eléctricamente grandes y complejos. En [30], se desarrolla un método de óptica física iterativa en el dominio del tiempo acelerado para analizar la dispersión de objetivos eléctricamente grandes y complejos, y se realiza una IFFT para obtener la imagen ISAR en condiciones de ancho de banda y ángulo pequeños.

Con el objetivo de obtener imágenes ISAR para objetivos eléctricos grandes en condiciones de gran ancho de banda y gran angular, este artículo está dedicado a un esquema de imágenes ISAR combinando SBR acelerado por GPU basado en la aceleración de árbol GPU y BVH con BP acelerado por GPU.

algoritmo. Para mejorar la eficiencia de la intersección de rayos, se construye una estructura de árbol BVH

SBR acelerado basado en GPU y aceleración de árbol BVH con el algoritmo B acelerado por GPU. Para mejorar la eficiencia de la intersección de rayos, se construye una estructura de árbol BVH de acuerdo con la estructura objetivo, que se implementa en C++AMP para lograr el cálculo de la aceleración paralela de la GPU. El método SAH se incorpora a la división del bo delimitador de escena, mitigando efectivamente el impacto de la superposición del cuadro delimitador en la eficiencia transversal del rayo de la estructura del árbol BVH. Realizar eficientemente simulaciones de imágenes ISAR de acuerdo con la estructura objetivo, que se implementa en C++AMP para lograr GPU paralela tion, se desarrolla un algoritmo de imágenes de BP acelerado basado en GPU en virtud de un cálculo de aceleración Arquitectura de dispositivo unificada (CUDA).

informática. El método SAH se incorpora al cuadro delimitador de la escena. división, mitigando efectivamente el impacto de la superposición del cuadro delimitador en el recorrido del rayo utilizando la eficiencia de la estructura de árbol BVH. Para realizar eficientemente la simulación de imágenes ISAR, un El algoritmo de imágenes de BP acelerado basado en GPU se desarrolla en virtud de un cómputo unificado dispositivo (CUDA).

Este artículo está organizado de la siguiente manera: La Sección 2 presenta un SBR acelerado por GPU. Estructura de árbol BVH. En la Sección 3, se presenta el algoritmo de imágenes de BP acelerado por GPU. En la Sección 4, se presentan los resultados y la discusión, y se realizan varias simulaciones por arquitectura de formado para confirmar la viabilidad y eficiencia de nuestro esquema combinando aceleración por GPU SBR combinado con el algoritmo BP para una simulación rápida de imágenes ISAR en gran angular. condiciones de ancho de banda amplio. La sección 5 concluye este artículo.

Este artículo está organizado de la siguiente manera: La Sección 2 presenta un SBR acelerado por GPU usando Estructura de árbol BVH. En la Sección 3, se presenta el algoritmo de imágenes de BP acelerado por GPU. En la Sección 4, se presentan los resultados y la discusión, y se realizan varias simulaciones para confirmar la viabilidad y eficiencia de nuestro esquema combinando aceleración por GPU SBR con el algoritmo BP para una simulación rápida de imágenes ISAR en gran angular y gran angular 2. Un condiciones de ancho de banda. La sección 5 concluye este artículo.

2. Un SBR acelerado por GPU que utiliza la estructura de árbol BVH

2.1. Cálculo de dispersión múltiple utilizando PO y GO

Como método de aproximación de alta frecuencia, el método de disparo y rebote de rayos es una combinación de PO y GO, que utiliza GO para rastrear el reflejo de la onda electromagnética. Como método de aproximación de alta frecuencia, el método de disparo y rebote de rayos La ruta de dispersión y PO para calcular el campo de dispersión, lo que resulta en una gran ventaja en la resolución de problemas de dispersión electromagnética para objetivos complejos [31]. De acuerdo con la ruta de aproximación PO y PO para calcular el campo de dispersión, lo que resulta en una gran ventaja en resuelven problemas de dispersión electromagnética para objetivos complejos [31]. Según PO

En la teoría mate, las superficies del objetivo se dividen en áreas brillantes y oscuras, dependiendo de si se las superficies están iluminadas por ondas electromagnéticas o no. El campo total disperso según la teoría aproximada, se supone que el área oscura es cero. Sin embargo, esta suposición sólo es válida cuando se decide si las superficies La longitud de onda de la onda electromagnética es mucho más pequeña que la geometría del objetivo. Se supone que Resulta difícil determinar el campo total del área sobre el objetivo que no es geometría directa. Se vuelve difícil iluminado por la onda incidente cuando las dimensiones del objetivo son muy grandes a lo largo de th no está iluminado dirección perpendicular de la onda incidente. Suponiendo un campo total de cero en la sombra grande a lo largo de la superficie implicaría una discontinuidad en el campo en el límite sombreado. Por lo tanto, t en la superficie sombreada Para resolver la discontinuidad en el campo límite, se debe agregar una integral de línea. Por lo tanto, para resolver la límite [32,33]. En la Figura 1, se ilustra un diagrama esquemático de la dispersión múltiple de SBR en el tratado; en el que solo se ilustran la 1.ª reflexión (en flechas amarillas) y la 2.ª reflexión (en verde) , en la que sólo la 1.ª (flechas) están representadas.

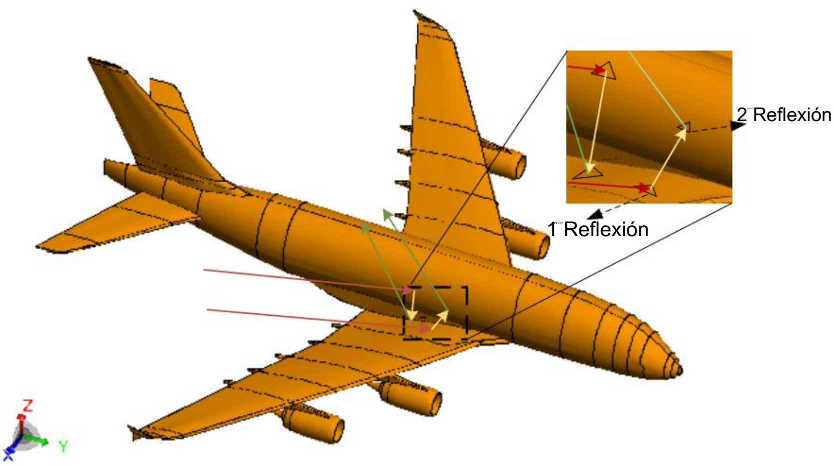


Figura 1. Diagrama esquemático de la dispersión múltiple.

El campo PO del objetivo del conductor eléctrico perfecto (PEC) en la posición sr se puede expresar como [34]

$$E^{s}(r_s) = -jk\eta \iint_S \hat{k}_s \times \hat{k}_s \times J\, G(r_s, r')\, dS' = -jk\eta \frac{e^{-jk(r_s - r')}}{2\pi r_s} \iint_S \hat{k}_s \times \hat{k}_s \times \hat{n} \times \hat{h}_i\, e^{-jk(\hat{k}_i - \hat{k}_s)\cdot r'} dS' \quad (1)$$

$$E_s(r_s) = -jk\eta \int_S \hat{k}_s \times \hat{k}_s \times J(r') G(r_s, r') dS' = -jk\eta \frac{e^{-jkr_s}}{2\pi r_s} \int_S \left\{ \hat{k}_s \times \hat{k}_s \times \left(\hat{n} \times \hat{h}_i \right) \right\} e^{-jk(\hat{k}_i - \hat{k}_s) \cdot r'} dS' \quad (1)$$

donde $\hat{h}_i$ es el vector de dirección del campo magnético incidente y $\hat{k}_i$ es el vector de dirección de propagación de ondas electromagnéticas. $\hat{k}_s$ es el vector unitario de la dirección de la onda de dispersión, r' es la ubicación del punto fuente, y S representa el área iluminada por la onda incidente. En la zona de campo lejano, la función de Green en el espacio libre se puede aproximar como [35]

$$G(r, r') \approx \frac{e^{-jk|r - r'|}}{4\pi|r|} \approx \frac{e^{-jk(R - r')}}{4\pi r} \quad (2)$$

Cuando el radio de curvatura en un punto del objetivo es mucho mayor que la longitud de onda de la onda incidente, se puede aplicar la aproximación del plano tangencial y la corriente inducida se puede expresar como [36]

$$J = \hat{n} \times H = \begin{cases} 2\hat{n} \times H_i & \text{Región iluminada} \\ 0 & \text{Región de sombra} \end{cases} \quad (3)$$

En la Figura 2, D_m es el área del tubo de rayos en el rayo m, y D_{m+1} es el área del tubo en el rayo reflejado del rayo m. Los rayos provocan cambios en la amplitud y fase del campo eléctrico a medida que se propagan y reflejan, y se proporciona información sobre la intensidad y la fase del campo eléctrico de cada tubo de rayos en cada punto de reflexión. En GO, el campo eléctrico en r_m está relacionado con el campo eléctrico en su r reflejado según la siguiente ecuación.

$$E^s(r_{m+1}) = \Gamma_m \cdot E^s_m(r_m) \cdot (DF)_m \cdot e^{-j\beta} \quad (4)$$

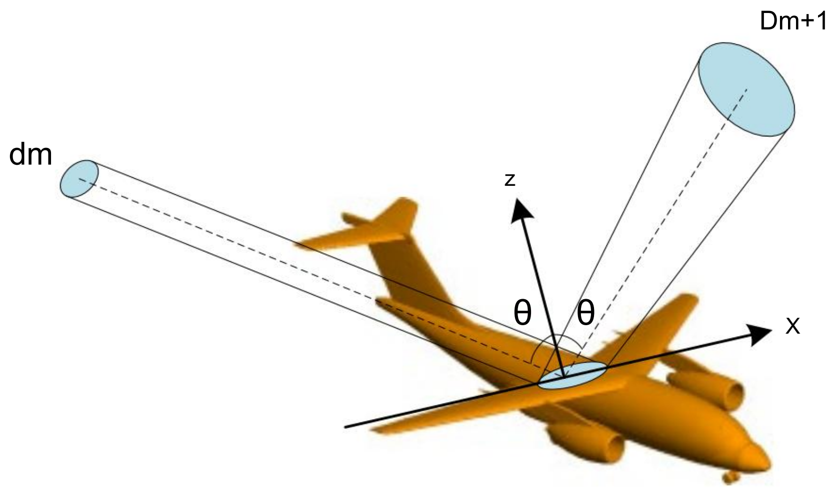


Figura 2. Diagrama esquemático de la reflectancia del haz. La forma del tubo de rayos cambia después de cada reflexión, que está determinada por el factor de divergencia $(DF)_m$.

En la ecuación (4), Γ_m denota el coeficiente de reflexión en r_m. Para un conductor perfecto, $\Gamma_m = -1$ para polarización horizontal y $\Gamma_m = 1$ para polarización vertical. El factor de divergencia $(DF)_m$ en r_m generalmente se denota por $(DF)_m \approx D_m/D_{m+1}$. $\beta = k_0|r_{m+1} - r_m|$, que representa la diferencia de fase entre dos puntos de reflexión vecinos. Sustituyendo la Ecuación (1) en la Ecuación (4), el campo disperso después de la reflexión del rayo m puede ser expresado como [37–40]

$$E^s_{m+1}(r_m) = -jk\eta\Gamma_m(DF)_m e^{-jk|r_{m+1} - r_m|} \sum_{x\Delta x} \hat{k}_s \times \hat{k}_s \times JG(r_m, r')dS = -jk\eta\Gamma_m(DF)_m e^{-jk|r_{m+1} - r_m|} \frac{e^{-jk(r_m - r)}}{2\pi|r|} \sum_{x\Delta x} \hat{k}_s \times \hat{k}_s \times \hat{n} \times \hat{h}_i e^{-jk(\hat{k}_i - \hat{k}_s) \cdot r'} dS' \quad (5)$$

En la ecuación (5), $\sum_{x} mi^{\overline{\Delta x}^{correcto}}(rm)$ es el campo PO generado por el elemento de faceta golpeado por el m-ésimo rayo, y x es el número de elementos facetarios golpeados. El campo disperso resultante para cada rayo se superpone posteriormente para obtener el campo disperso total.

$$mi_{s}^{SBRtotal}(rm) = \sum_{1}^{yo} \sum_{x} mi^{\overline{\Delta x}^{correcto}}(rm) \quad (6)$$

En la ecuación (6), I es el número total de rayos y Δx es el área de algún elemento de la cara triangular que fue golpeado.

2.2. Trazado de rayos acelerado por GPU utilizando la estructura de árbol BVH

2.2.1. Proceso de aceleración de GPU

En este artículo, el SBR acelerado por GPU se paraleliza utilizando el paralelismo masivo acelerado por C++ (C++AMP). En comparación con las CPU, las GPU poseen una mayor cantidad de núcleos, lo que las hace más adecuadas para el procesamiento paralelo masivo. C++AMP es una plataforma informática paralela heterogénea basada en C++ lanzada por Microsoft, que es un modelo de programación nativo con la ventaja de ejecutarse en dispositivos en la plataforma Windows [41]. La mayoría de los métodos de programación para GPU, como Direct Compute y OpenCL, requieren diferentes lenguajes de programación y compiladores. C++AMP unifica el lenguaje de programación y el compilador, lo que lo diferencia de otros enfoques. La biblioteca C++ AMP permite el cálculo paralelo a través de un conjunto de abstracciones y una API de alto nivel, accediendo directamente al hardware GPU subyacente a través de Direct Compute [42,43]. Es importante asignar una matriz para aplicar C++AMP para implementar la computación paralela. La plantilla de matriz está en el espacio de nombres de concurrencia. Se necesitan dos parámetros: uno para el tipo de elemento de colección y el otro para la dimensión. La dimensión de la matriz se establece según el tipo de elementos de la colección en este método de papel.

Por ejemplo, al recopilar datos de frecuencia de barrido y definir una <matriz a (recuentos de frecuencia)>, este ejemplo define una matriz unidimensional cuyo tamaño es el número de frecuencias. Las matrices desempeñan un papel extremadamente importante en C++ AMP al representar una vista que puede acceder a datos en la GPU y encapsular matrices o vectores de C++, que son matrices en <accelerator_view>. En C++AMP, la GPU no es el único acelerador y cada acelerador tiene su propia vista predeterminada. Una vez que se haya construido la matriz, los datos se transferirán a la memoria de la GPU, donde la GPU podrá acceder a ellos directamente. La función <parallel_for_each> se utiliza para ejecutar tareas de cálculo paralelas. La función <parallel_for_each> es una función de ejecución paralela en C++AMP que acepta un rango de índices y una función lambda, y ejecuta esta función lambda en paralelo en la GPU para cada índice. La función <parallel_for_each> delega tareas de computación paralela a las funciones del núcleo de la GPU , que se pueden asignar directamente al hardware de la GPU a través de Direct Compute API.

La palabra clave <restrict(amp)> se utiliza para especificar que la función se ejecutará solo en la GPU. Se utiliza para identificar bloques específicos de código y funciones lambda que se ejecutarán en la GPU y permite al compilador optimizar la función para instrucción única, subprocesos múltiples (SIMT) [44]. Las operaciones dentro de la función utilizarán instrucciones SIMT para lograr el efecto de cálculo paralelo de múltiples subprocesos de una sola instrucción. C++AMP sincroniza los datos y los copia desde la memoria de la GPU a la memoria del host después de que la función <parallel_for_each> ejecuta la tarea de cálculo paralelo.

La Figura 3 muestra un proceso de cálculo paralelo para C++AMP utilizando la API de Computación Directa para enviar instrucciones paralelas al dispositivo GPU. La estructura de árbol BVH, así como los datos de rayos, etc., se construyen en la CPU y se almacenan en la memoria global de la GI Los datos se copiarán automáticamente desde el host de la CPU a la memoria de la GPU mediante la creación de la matriz <array_view>, lo que permitirá que la GPU acceda a estos datos directamente. La memoria constante de la GPU almacenará datos que permanecerán sin cambios durante el cálculo paralelo. Texture Memory se utiliza para almacenar datos durante la renderización del modelo. Hay muchos multiprocesadores de transmisión (SM) en la arquitectura de hardware de la GPU, y los SM en las GPU utilizan la arquitectura SIMT. Cada SM contiene varios procesadores de transmisión (SP) y cada SP

ing la matriz <array_view> , lo que permite que la GPU acceda a estos datos directamente. La memoria constante de la GPU almacenará datos que permanecerán sin cambios durante el cálculo paralelo.

Texture Memory se utiliza para almacenar datos durante la renderización del modelo. Hay muchos multiprocesadores de transmisión (SM) en la arquitectura de hardware de la GPU, y los SM en las GPU utilizan la arquitectura SIMT. Cada SM contiene una cantidad de procesadores de transmisión (SP) y cada SP corresponde a un subproceso. Un solo warp se compone de 32 subprocesos, y la arquitectura de la GPU determina el número de subprocesos por warp. Warp es la programación y la ejecución corresponde a un hilo. Una sola urdimbre se compone de 32 hilos, siendo el número de unidad de ción en SM, donde los hilos en cada urdimbre ejecutan cálculos en paralelo. Como antes, los subprocesos por deformación están determinados por la arquitectura de la GPU. Warp es la programación y Dicho claramente, SM unión <parallel_for_each> asigna tareas directamente a la unidad de ejecución de GPU subyacente en SM, donde los sub_r_c_esos en cada warp ejecutar cálculos en paralelo. Como hardware a través de la API Direct Compute. La GPU asigna instrucciones a las deformaciones en el hardware de la Como se indicó anteriormente, la función <parallel_for_each> asigna tareas directamente al sistema subyacente SM con arquitectura SIMT. La arquitectura SIMT permite los subprocesos dentro de cada uno en el SM con la GPU a través de la API Direct Compute. La GPU asigna instrucciones a los warps warp para ejecutar una secuencia de instrucciones, incluida la detección de colisiones, trazado de rayos, PO cada warp arquitectura SIMT. La arquitectura SIMT permite que los hilos dentro para ejecutar una secuencia de instrucciones, incluida la detección de colisiones, trazado de rayos, cálculo de campo y otras operaciones, en modos paralelos. En la arquitectura SIMD, el cálculo del campo PO y otras operaciones, en modos paralelos. En la arquitectura SIMD, el La GPU asigna un hilo a cada rayo. La estructura de árbol BVH se obtiene accediendo al hilo de la GPU y asignando un hilo a cada rayo. La estructura del árbol BVH se obtiene accediendo al hilo. la memoria global de la GPU y es responsable de calcular la ruta de trazado de rayos y la memoria global de la GPU y es responsable de calcular la ruta de trazado de rayos y el PO Campo de orden de compra. En la función <parallel_for_each> , detección de intersecciones y campo de aceleración paralela . En la función <parallel_for_each> , detección de intersecciones y aceleración paralela La unión de rayos y cuadros delimitadores se logra atravesando la estructura del árbol BVH. Después de los rayos y los cuadros delimitadores se logra atravesando la estructura del árbol BVH. Después de la se completa la ejecución de la función <parallel_for_each> , C++AMP copia el scat- se completa la ejecución de la datos de campo almacenados desde la memoria de la GPU a la memoria del host. La CPU sumará los campos dispersos de función <parallel_for_each> , C++AMP copia el scat- datos de campo desde la memoria de la GPU a la memoria del host. La CPU sumará los campos dispersos de todos los rayos para obtener el campo de dispersión total del objetivo. de todos los rayos para obtener el campo de dispersión total del objetivo.

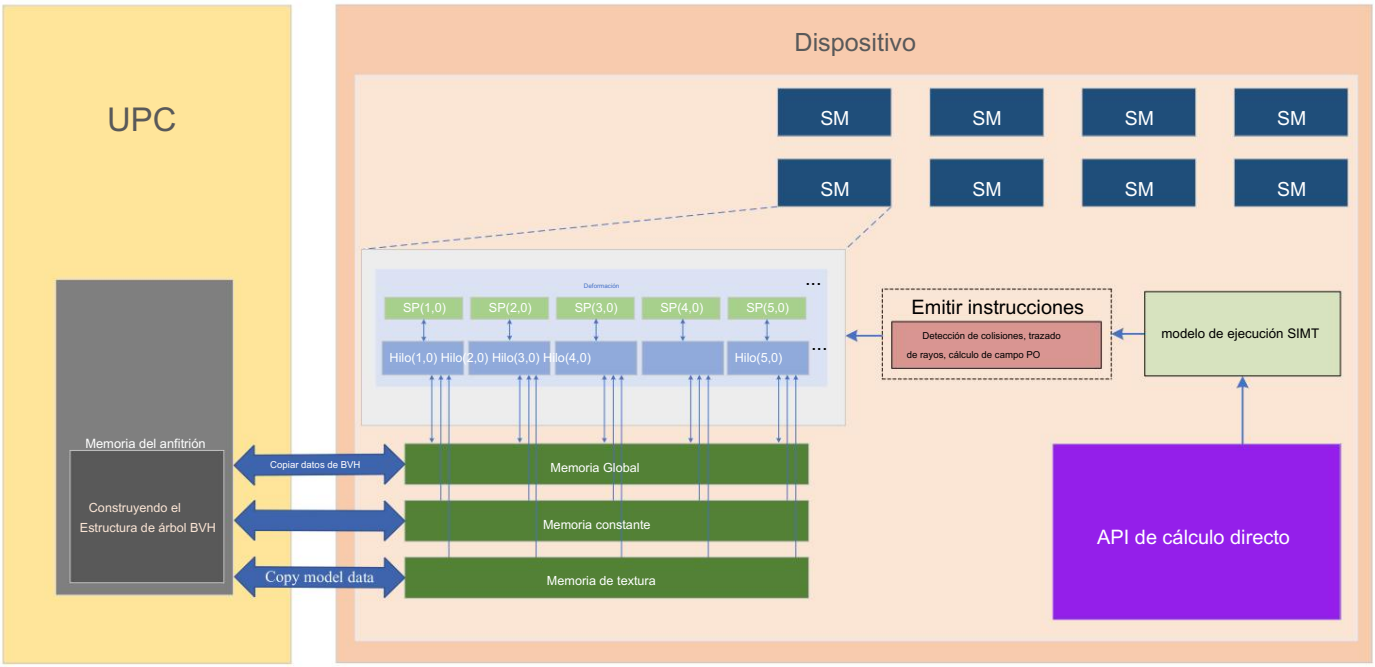


Figura 3. El proceso de cálculo paralelo de C++AMP utilizando la API Direct Compute para enviar en paralelo instrucciones al dispositivo (GPU).

Figura 3. El proceso de cálculo paralelo de C++AMP utilizando la API Direct Compute para enviar en paralelo instrucciones al dispositivo (GPU).

La información del rayo se inicializa y los nodos, así como la estructura del árbol BVH, se generan en la CPU como se ilustra en la Figura 4. La CPU compartirá los datos del árbol BVH con la memoria global de la GPU y compartirá estos datos con cada subproceso.

El árbol se genera en la CPU como se ilustra en la Figura 4. La CPU compartirá los datos del Árbol BVH a la memoria global de la GPU y compartirá estos datos con cada hilo. En la arquitectura SIMT , la API Direct Compute emite instrucciones al warp.

Arquitectura SIMT, las instrucciones se emiten al warp mediante la API Direct Compute.

Los hilos en una urdimbre ejecutarán las instrucciones recibidas de forma secuencial y en paralelo, y cada hilo en una urdimbre será responsable del cálculo de un rayo. En esta etapa, todos

Los hilos en una urdimbre ejecutarán las instrucciones recibidas de forma secuencial y en paralelo, y cada hilo en una cada hilo en una urdimbre será responsable del cálculo de un rayo. En esta etapa, se registran todos los rayos que atraviesan el árbol BVH que cruza el objetivo y sus rayos reflejados.

Se registren los rayos que atraviesan el árbol BVH que cruza el objetivo y se rastrean sus rayos reflejados hasta que el rayo abandona la superficie del objetivo. La CPU calcula el campo disperso de

se rastrean hasta que el rayo abandona la superficie objetivo. La GPU calcula el campo disperso de este rayo y transfiere los datos a la CPU. Finalmente, la CPU combina el campo disperso.

este rayo y transfiere los datos a la CPU. Finalmente, la CPU combina los datos del campo disperso de todos los rayos para determinar el campo de dispersión total del objetivo.

datos de todos los rayos para determinar al campo de dispersión total del objetivo.

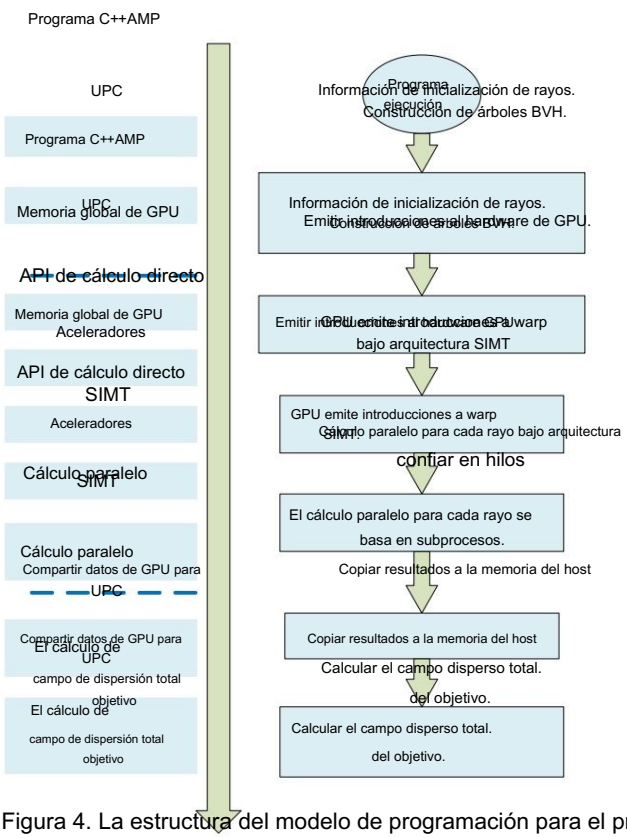


Figura 4. La estructura del modelo de programación para el proceso de aceleración de GPU.

Figura 4. La estructura del modelo de programación para el proceso de aceleración de GPU.

Figura 4. La estructura del modelo de programación para el proceso de aceleración de GPU.

2.2.2. Algoritmo de trazado de rayos utilizando el árbol BVH

2.2.2. Algoritmo de trazado de rayos utilizando el árbol BVH

La Figura 5 ilustra el proceso de trazado de rayos usando la estructura de árbol BVH con el multip. La dispersión múltiple de una onda electromagnética entre las estructuras de árbol. GO se utiliza en el proceso de rastrear la ruta de dispersión de la onda electromagnética entre la superficie triangular y rastrear la ruta de dispersión de la onda electromagnética entre las tuplas de la superficie triangular. y PO se usa para calcular el campo disperso de la onda electromagnética cuando y PO se usa para calcular el campo disperso de la onda electromagnética cuando golpea las tuplas de superficie triangular.

Figura 5 ilustra el proceso de trazado de rayos usando la estructura de árbol BVH con el multip. La dispersión múltiple de una onda electromagnética entre las estructuras de árbol. GO se utiliza en el proceso de rastrear la ruta de dispersión de la onda electromagnética entre la superficie triangular y rastrear la ruta de dispersión de la onda electromagnética entre las tuplas de la superficie triangular. y PO se usa para calcular el campo disperso de la onda electromagnética cuando y PO se usa para calcular el campo disperso de la onda electromagnética cuando golpea las tuplas de superficie triangular.

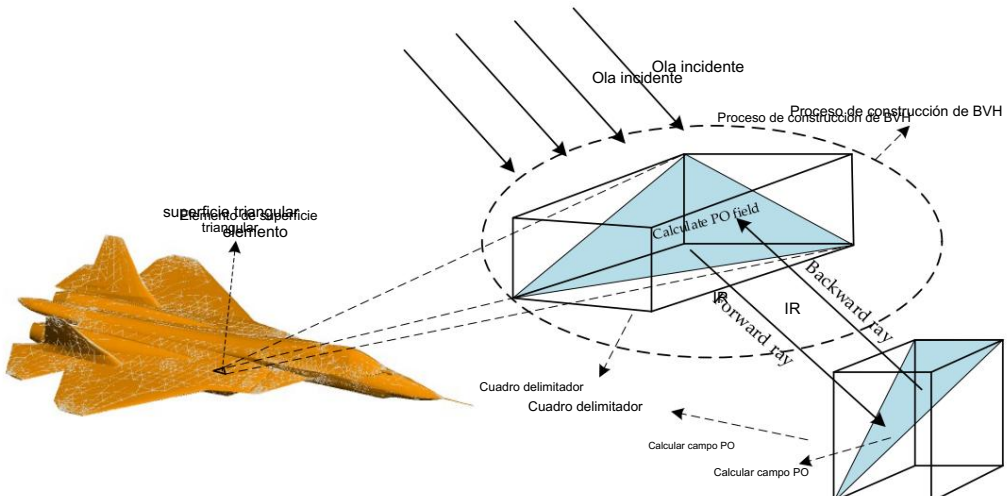


Figura 5. Estructura de árbol BVH que representa la dispersión múltiple entre parches de superficie triangulares.

Figura 5. Estructura de árbol BVH que representa la dispersión múltiple entre parches de superficie triangulares.

Para resolver el lento proceso de intersección de rayos en el trazado de rayos, recurrimos Para resolver el lento proceso de intersección de rayos en el trazado de rayos, recurrimos a la estructura del árbol BVH sobre la base de la aceleración de GPU. Un árbol BVH es un gráfico por computadora. Para resolver el lento proceso de intersección de rayos en el trazado de rayos, recurrimos a la estructura del árbol BVH sobre la base de la aceleración de GPU. Un árbol BVH es un gráfico por computadora. que es una técnica de detección de intersecciones de rayos basada en tuplas. Un árbol BVH se divide en una estructura de árbol BVH según la aceleración de la GPU. Un árbol BVH es una estructura gráfica de las tuplas en una estructura jerárquica de conjuntos disjuntos, que se usa ampliamente en la estructura de trazado de rayos, que es una técnica de detección de intersecciones de rayos basada en tuplas. Un árbol BVH divide las tuplas en una estructura jerárquica de conjuntos disjuntos, que se usa ampliamente en el trazado de rayos y detección de colisiones. Construiremos el cuadro delimitador que encierra el objetivo según las tuplas en una estructura jerárquica de conjuntos disjuntos, que se usa ampliamente en trazado de rayos y detección de colisiones

Construiremos el cuadro delimitador que encierra el objetivo de acuerdo

a las características geométricas objetivo, y su estructura de caja envolvente es una caja delimitadora axisimétrica, una caja delimitadora AABB. En la estructura de árbol de BVH, todas las tuplas se almacenan en t y detección de colisiones. Construiremos el cuadro delimitador que encierra el objetivo de acuerdo con las características geométricas objetivo y su estructura de caja envolvente son un límite simétrico de eje. Los nodos de hoja del árbol BVH y los nodos del medio almacenan la información del cuadro. Finalmente, t un cuadro delimitador AABB. En la estructura de árbol BVH, todas las tuplas se almacenan en la hoja. La información de toda la escena se almacena en la estructura de árbol de BVH [45]. Al atravesar el cuadro t, del medio almacenan la información del cuadro. Finalmente, el conjunto Árbol BVH, el rayo primero juzgará si se cruza con la caja o no. Si no son entrenodos del árbol BVH, los nodos información de la escena se almacene en la estructura del árbol BVH [45]. Al atravesar el árbol BVH seccionado, entonces omitirá todas las tuplas en el cuadro, logrando una mejora para que la eficiencia de rastreo. el rayo juzgará primero si se cruza con la caja o no. Si no se cruza, entonces las tuplas del cuadro se dividen y se omitirán todas

En la Figura 6, la escena contiene ocho tuplas. En el Paso 1, las tuplas se dividen en dos partes en la escena. Si los rayos de luz no se cruzan con el cuadro delimitador, entonces las partes de la escena. Si los rayos de luz no se cruzan con el cuadro delimitador, entonces no se cruzará con las tuplas, lo que puede excluir la mitad de las tuplas a la vez y no se cruzará con las tuplas, lo que puede excluir la mitad de las tuplas a la vez y, por lo tanto, Reducir el número de intersecciones. En el Paso 2, continuamos dividiendo las tuplas en hal para reducir el número de intersecciones. En el Paso 2', continuamos dividiendo las tuplas por la mitad y y al hacerlo, la complejidad de la intersección de rayos se puede reducir de O n( ) t al hacerlo, la complejidad de la intersección de rayos se puede reducir de O(n) a O(log(n)) [46]. Después de la división recursiva por mitad, el cuadro delimitador que rodea solo se puede encontrar una tupla. Si un rayo cruza este cuadro delimitador, el rayo continúa en tupla. La puede ser encontrado. Si un rayo intersecta este cuadro delimitador, el rayo continúa intersecándose con el ecuación del parámetro del rayo se puede expresar como se cruza con la tupla. La ecuación del parámetro del rayo se puede expresar como

$$r(t) = o + t \cdot d \tag{7}$$

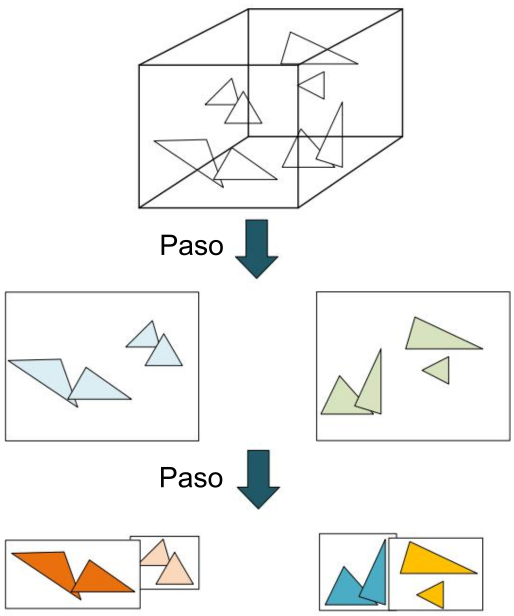


Figura 6. Proceso de segmentación de tuplas de escenarios.

En la ecuación (7), o y d son el punto de inicio del rayo y el vector de dirección del rayo normalizado, respectivamente. Sustituyendo la ecuación del parámetro del rayo en la ecuación del plano implícito de el plano donde se encuentra la tupla, la ecuación del plano implícito se puede escribir como

$$N^t (o + t \cdot d) = c \tag{8}$$

De la ecuación (8), el parámetro t correspondiente al punto de intersección del rayo puede resolver como

$$t = \frac{c - N^T \cdot o}{N^T \cdot d} \tag{9}$$

En este artículo, el elemento de superficie es un elemento de superficie triangular y el plano la ecuación paramétrica del elemento de superficie triangular es

$$f(u, v) = (1 - u - v) \cdot p0 + u \cdot p1 + v \cdot p2 \tag{10}$$

Electrónica 2024, 13, x PARA REVISIÓN POR PARES 9 de 24

donde u y v son las coordenadas del centro de masa del triángulo, que satisfacen u ≥ 0 y u+v ≤ 1 . El elemento de cara triangular puede considerarse como el mapeo de la unidad donde u y v son las coordenadas del centro de masa del triángulo, satisfaciendo u ≥ 0 elemento de cara triangular en sus tres bordes.

y u+v ≤ 1 . El elemento de cara triangular puede considerarse como el mapeo de la unidad donde u y v son elemento de cara triangular en sus tres aristas

(11)

u + v ≤ 1. El elemento de cara triangular puede considerarse como el mapeo del triángulo unitario elemento de cara en sus tres bordes

Combinando la ecuación (11) con la ecuación (7), se puede obtener

(11)

Combinando la ecuación (11) con la ecuación (7), se puede obtener

(12)

(12)

En la ecuación (12), M −es la matriz que transforma un elemento de cara triangular en un elemento de cara triangular unitario en el plano u,v , en el que el rayo mapeado es ortogonal a En la ecuación (12), M −es la matriz que transforma un elemento de cara triangular en un elemento El elemento de cara triangular unitario en el plano u, v , en cual el rayo mapeado es ortogonal al plano se ilustra en la Figura 7. La Figura 7a es el caso antes del mapeo, y la Figura 7b es el elemento de cara triangular unitario. El mapeo de elementos de superficie triangulares en u,v El plano se ilustra en la Figura 7. La Figura 7a es el caso antes del mapeo, y la Figura 7b es el caso correspondiente a la Ecuación (11). La figura 7c representa la intersección del rayo con el elemento de superficie triangular unitario después del mapeo.

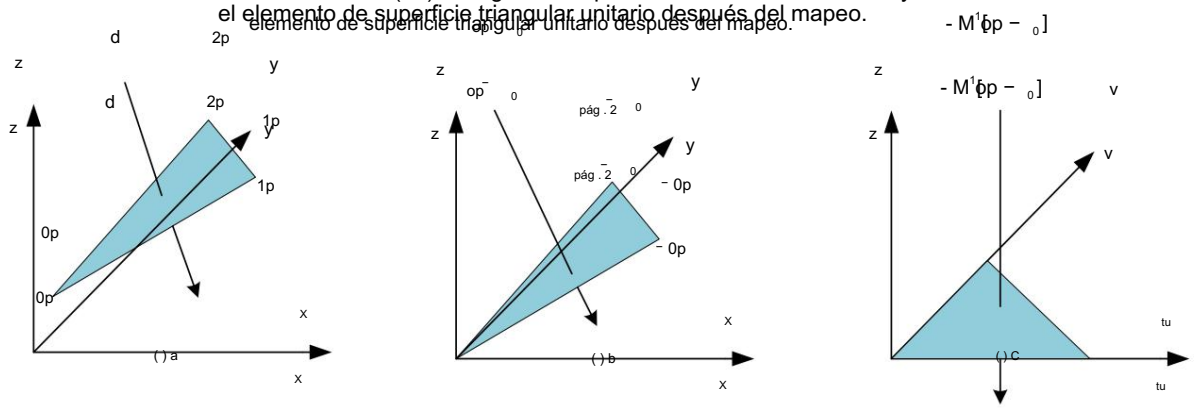


Figura 7. Mapeo de elementos de superficie triangulares en el plano u,v . a) El caso previo al mapeo; (b) El caso correspondiente a la Ecuación (11); (c) La intersección del rayo con la unidad triangular Figura 7. Mapeo de elementos de superficie triangulares en el plano u, v . a) El caso previo al mapeo; (b) El caso correspondiente a la Ecuación (11); (c) La intersección del rayo con el elemento triangular unitario después del mapeo. elemento de superficie después del mapeo.

Según el método de media división, los cuadros delimitadores de la escena pueden superponerse o intersecarse, y la superposición de cuadros delimitadores se ilustra en la Figura 8. En la Figura 8, según el método de media división, los cuadros delimitadores en la escena pueden superponerse cuando se cruce, y la superposición de los cuadros delimitadores se ilustra en la Figura 8. En la Figura 8, cuando la luz atraviesa el árbol BVH, la superposición de los cuadros delimitadores hace que la luz cruce para cruzar ambos cuadros delimitadores, lo que conduce a una disminución en la eficiencia de la intersección.

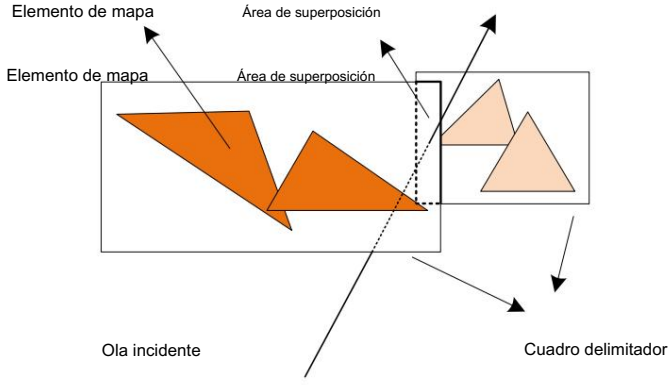


Figura 8. Diagrama de superposición de los cuadros delimitadores de los nodos secundarios.

Figura 8. Diagrama de superposición de los cuadros delimitadores de los nodos secundarios.

Para eliminar el fenómeno de superposición de cajas de sobre, se adopta el método de división SAH en lugar del método de media división en el proceso de configuración del árbol BVH. Para eliminar el fenómeno de superposición de cajas de sobre, se adopta la estructura de división SAH. El SAH se basa en el método de división heurística de área de superficie, y en el proceso de construcción del árbol BVH se adopta el método posterior en lugar del método de media división. El SAH se basa en el método de división heurística de área de superficie, y después

Para eliminar el fenómeno de superposición de cajas envolventes, en el proceso de construcción del árbol BVH se adopta el método de división SAH en lugar del método de media división . El SAH se basa en el método de división heurística del área de superficie, y después de sumar el SAH, podemos estimar la probabilidad de que el rayo de luz incida en las cajas envolventes en términos del tamaño del área de superficie de la caja envolvente principal en la que hay dos o más cajas envolventes infantiles superpuestas.

En la estructura de árbol de BVH, bajo el supuesto de que el nodo actual tiene tres cuadros delimitadores A, B y C, el costo de cruzar el rayo con el nodo actual es

$$c(A, B, C)= p(A)\sum_{i\ A} t(i)+p(B)\sum_{i\ B} t(j)+p(C)\sum_{i\ C} t(k) + ttrav \quad (13)$$

En la ecuación (13), ∑ t(i) es el costo de intersección de cada subcaja, y t(i) es la i-ésima tupla en la caja secundaria. p(A), p(B) y p(C) son las probabilidades de que la luz incida en los objetos en los cuadros delimitadores A, B y C, respectivamente. ttrav es el costo de la luz que atraviesa el árbol BVH.

En SAH, utilizamos el área de superficie del cuadro delimitador secundario en el nodo principal en lugar de la probabilidad de que el rayo golpee el cuadro delimitador. Suponiendo que las áreas de superficie de los cuadros delimitadores secundarios A, B y C son S(A), S(B) y S(C), respectivamente, y que la superficie del cuadro delimitador del nodo principal D es S(D) , La ecuación (13) se puede reescribir como

$$c(A, B, C) = \frac{S(A)}{\text{DAKOTA DEL SUR)}}\sum_{i\ A} t(i) + \frac{S(B)}{\text{DAKOTA DEL SUR)}}\sum_{i\ B} t(j) + \frac{S(C)}{\text{DAKOTA DEL SUR)}}\sum_{i\ C} t(k) + ttrav \quad (14)$$

Después de resolver el método de división óptimo calculando el valor mínimo de la ecuación (14), el recorrido de rayos del árbol BVH es más eficiente [47].

3. Algoritmo de imágenes de BP acelerado por GPU

3.1. Algoritmo BP para imágenes ISAR

En condiciones de ángulo pequeño y ancho de banda pequeño, las imágenes ISAR se pueden obtener aproximadamente realizando la transformada inversa de Fourier de datos de campo retrodispersados 2D, y las imágenes ISAR resultantes se componen de los centros de dispersión del objetivo con su coeficiente de reflexión electromagnética. Debido a su idoneidad para el procesamiento paralelo de GPU y su capacidad para obtener imágenes ISAR en cualquier modo, el algoritmo BP y sus versiones modificadas se utilizan ampliamente en aplicaciones de imágenes SAR/ISAR. El algoritmo de imágenes de BP es un método con alta precisión de imágenes. Sin embargo, debido a su alta complejidad, el algoritmo BP no es tan bueno como otros algoritmos de imágenes en términos de velocidad de obtención de imágenes. Con el objetivo de obtener imágenes ISAR para objetivos eléctricos grandes en condiciones de gran ancho de banda y gran angular, en este artículo se desarrolla un algoritmo de imágenes de BP acelerado basado en GPU en virtud de CUDA, manteniendo al mismo tiempo la precisión de las imágenes del algoritmo de BP.

La idea fundamental del algoritmo BP implica superponer coherentemente los ecos calculados de cada pulso mediante la transmisión de pulsos electromagnéticos y calcular el retardo de tiempo bidireccional entre los puntos de píxeles en el área de imagen y el radar en el momento de cada pulso. La superposición depende de la relación de fase entre los puntos de píxeles. Si los ecos están en fase, los ecos de los puntos de píxeles superpuestos se vuelven cada vez más fuertes. Cuando se superponen puntos de píxeles con diferentes fases, el efecto es más débil. Como algoritmo preciso en el dominio del tiempo, el perfil de rango en el algoritmo BP se obtiene utilizando la técnica de compresión de pulso, similar al algoritmo Range-Doppler.

El procesamiento de la dirección azimutal se logra calculando los ecos de los puntos de píxel para una superposición coherente, que está relacionada con el ángulo de rotación del objetivo con respecto al radar. Se observa que la resolución azimutal aumenta a medida que aumenta el ángulo entre el objetivo y el radar, sin límite aparente. Para cualquier trayectoria de movimiento del objetivo, si la trayectoria de movimiento se puede predecir de antemano, entonces el algoritmo BP puede lograr imágenes precisas.

Figura 9 ilustra el diagrama esquemático de proyección hacia atrás para imágenes ISAR. xyz es un sistema de coordenadas global. El sistema de coordenadas local XY es la cuadrícula para el área de imágenes. Ra y Rb denotan la distancia del radar al objetivo en momentos a y b, respectivamente. El área de imágenes se divide en cuadrículas N × N. (Xa,Ya) y (Xb ,Yb) son las posiciones del objetivo en la cuadrícula de imágenes en los momentos a y b, respectivamente. v es la velocidad y dirección del objetivo en movimiento. La señal de transmisión es [48]

$$s(t, m) = \text{recto}\left(\frac{t}{t_p}\right) \cdot \exp\left(j2\pi f_0 t + jK\pi t^2\right) \quad (15)$$

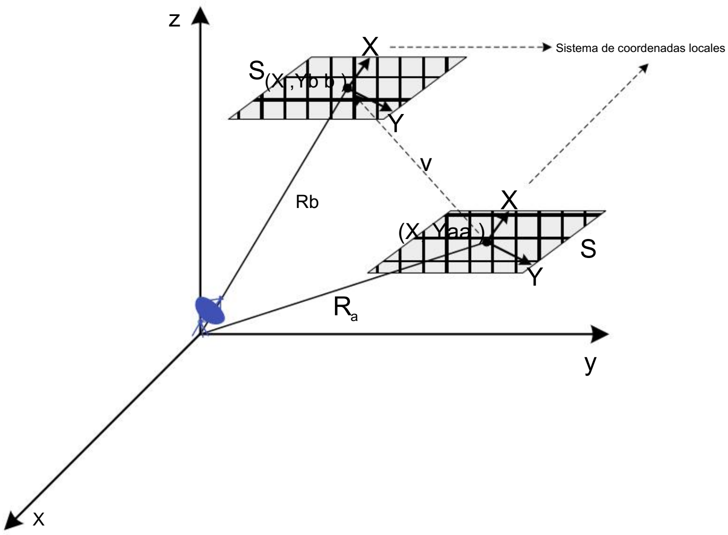


Figura 9. Diagrama esquemático de proyección hacia atrás para imágenes ISAR.

En la ecuación (15), t es el tiempo rápido. t_p es el ancho del pulso de la señal. m es tiempo lento. recto es una función de ventana rectangular. f_0 es la frecuencia portadora de la señal. $k = \frac{B}{t_p}$, K es la frecuencia de modulación de la señal. B es el ancho de banda de la señal. El eco recibido es el siguiente

$$sr(t, m) = \sum \text{recto}\left(\frac{t - \tau_{i,j}(m)}{t_p}\right) \exp\left(j2\pi f_0 \left(t - \tau_{i,j}(m)\right) + j\pi K \left(t - \tau_{i,j}(m)\right)^2\right) \quad (16)$$

En la ecuación (16), $\tau_{i,j}(m) = \frac{2\sqrt{(x(m)+X_i)^2+(y(m)+Y_i)^2+z(m)^2}}{c}$ es el retraso bidireccional del píxel (X_i, Y_i) en el plano de imagen al radar en el tiempo lento m. x(m), y(m) y z(m) son las posiciones del píxel (X_i, Y_i) en la cuadrícula de imágenes en el sistema de coordenadas del objetivo espacial. El filtro coincidente es el siguiente

$$h(t, m) = \exp\left(j2\pi f_0 t - jK\pi t^2\right) \quad (17)$$

En la ecuación (17), t_0 es el retardo bidireccional en la distancia más cercana entre el objetivo y el radar. Al convertir la convolución en el dominio del tiempo a la multiplicación en el dominio de la frecuencia, la salida del filtro coincidente se puede expresar como sigue

$$s_{out}(t, m) = IFFT\left(FFT\left(s(t, m)\right) \cdot FFT\left(h(t, m)\right)\right) \quad (18)$$

La alta resolución en azimutal se obtiene mediante la acumulación coherente de pulsos. La fórmula integral para la acumulación coherente es la siguiente [49]

$$Y_0(x, y) = \int_m s_{out}(t, m) \cdot \exp\left(j2\pi f_0 \tau_{i,j}(m)\right) dm \quad (19)$$

La energía de cada punto de píxel en la cuadrícula se superpone coherentemente sobre el tiempo de movimiento objetivo. El valor de píxel es acumulado por cada punto de píxel durante el tiempo de movimiento para sintetizar la imagen final.

3.2. Aceleración por GPU del algoritmo BP para imágenes

ISAR Lanzado por NVIDIA en 2006, CUDA es una plataforma informática paralela de propósito general y un modelo de programación basado en GPU. Los cálculos para tareas complejas se pueden realizar de manera más eficiente con la programación CUDA. En los últimos años, las técnicas de programación CUDA se han desarrollado tanto en hardware como en software [50]. En el momento de su lanzamiento inicial, CUDA era capaz de utilizar GPU con un número limitado de núcleos, normalmente en el rango de unas pocas docenas o unos cientos. En consecuencia, no fue posible hacer comparaciones significativas en términos de potencia computacional entre estas primeras GPU y las GPU disponibles en la actualidad [51]. Por ejemplo, NVIDIA RTX 2080 de NVIDIA, que se lanzó el 20 de septiembre de 2018, es una GPU con 2944 núcleos CUDA y una potencia de cálculo FP32 de 10,07 billones de veces por segundo. Sin embargo, la NVIDIA RTX 4090 ahora tiene 16.384 núcleos CUDA, con una potencia informática FP32 que alcanza los 82,58 billones de veces por segundo. En los últimos años, CUDA ha utilizado ampliamente sus poderosas capacidades de computación paralela en el campo de la computación científica [52].

En este artículo, realizamos el algoritmo de imágenes de PA altamente paralelo de CUDA. Los datos de campo dispersos obtenidos mediante el cálculo de SBR se cargan primero en la memoria global de la GPU desde la memoria del host bajo la función <cudaMemcpy>. Estos datos contienen información como azimut, ángulo de incidencia, frecuencia, puntos de muestreo de frecuencia, puntos de muestreo de ángulo, modo de polarización, etc. El espacio de memoria especificado se asigna para estos parámetros desde la GPU con un tamaño de memoria Nf × Nphi × Nf ft por el Función <cudaMalloc> . Las variables globales en el dispositivo se definen a través del símbolo _device_, incluida la señal de compresión de distancia y las variables utilizadas para almacenar los datos de campo dispersos sin procesar, las variables utilizadas para almacenar las coordenadas y la posición del radar, y las variables utilizadas para almacenar los resultados finales de las imágenes. . La compresión de pulsos de rango de los datos de eco sin procesar se realiza utilizando la biblioteca de funciones <cufft> en CUDA, mediante la cual se pueden realizar FFT e IFFT altamente paralelas . La señal de compresión de rango se copia en la memoria de la GPU utilizando la función <cudaMemcpy> con un tamaño de memoria asignado Nf × Nphi × Nf

El cálculo acelerado en paralelo se implementa principalmente en el dispositivo mediante la función del kernel CUDA. El kernel es un concepto importante en CUDA y es una función que se ejecuta en paralelo en un hilo del dispositivo. La función del núcleo se declara con el símbolo <_global_> y el número de subprocesos necesarios al llamar a esta función debe especificarse, específicamente mediante <<<grid, block>>>. La función del kernel es ejecutada por cada hilo. El proceso de enfoque azimutal del algoritmo BP se escribe como una función del núcleo y el número correspondiente de subprocesos se asigna a la función del núcleo para permitir el cálculo paralelo en la GPU.

En la Figura 10, la función kernel es responsable de calcular las distancias de todos los píxeles del radar en cada momento azimutal, así como la compensación de fase. La posición de la señal azimutal correspondiente está determinada por la distancia del punto de píxel al radar. Las imágenes ISAR que utilizan el algoritmo BP se obtienen superponiendo coherentemente los ecos de todos los puntos de píxeles en cada momento azimutal. Los bloques y subprocesos se definen como unidimensionales cuando se ejecuta la función CUDA del algoritmo BP. La distribución de hilos en cada bloque es،Ny, 1, 1 con Nx bloques en total y (Nx, 1, 1) distribución de bloques. Nx y Ny son el número de píxeles en la dirección x y el número de píxeles en la dirección y , respectivamente. Para optimizar la eficiencia de la función del núcleo en el procesamiento de datos a gran escala, las tareas computacionales para obtener imágenes de píxeles de más de 500 × 500 se calculan por lotes. El número máximo de puntos de píxeles calculados en cada lote es 500 × 500 para adaptarse a las limitaciones de recursos de la GPU. En la función del kernel, la GPU asigna 250.000 subprocesos para cada lote de tareas computacionales, siendo cada 500 subprocesos un bloque, para un total de 500 bloques. Cada hilo es responsable de calcular el valor de un punto de un píxel y se ejecutarán varios lotes de datos en p

Los puntos de píxeles calculados en cada lote son 500 500 × para adaptarse a las limitaciones de recursos de la GPU. En la función del kernel, la GPU asigna 250.000 subprocesos para cada lote de tareas computacionales, siendo cada 500 subprocesos un bloque, para un total de 500 bloques. Cada hilo es responsable de calcular el valor de un punto de un píxel y se ejecutarán varios lotes de datos en paralelo en el núcleo CUDA.

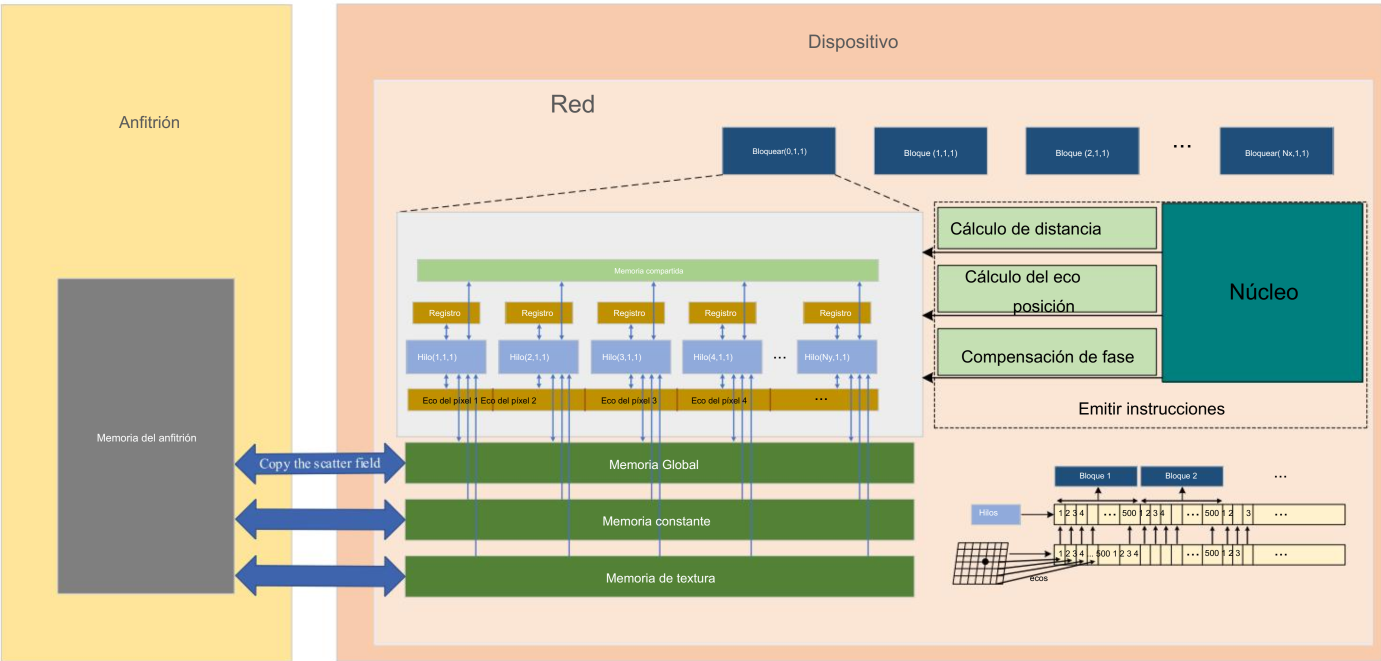


Figura 10. Proceso de cálculo paralelo del algoritmo BP para imágenes ISAR usando aceleración CUDA.

En la Figura 10, los datos del campo disperso se copian primero desde la memoria del host a la GPU. global de la GPU. Cada hilo puede acceder a los datos almacenados en la Memoria Global. La memoria constante se utiliza La memoria es una memoria de solo lectura que se utiliza para transmitir instrucciones enviadas por el host. La memoria se utiliza para almacenar datos de textura. El número de bloques y subprocesos está determinado por el número de píxeles Nx en la dirección x y el número de píxeles Ny en la dirección y di- rección, respectivamente. En un bloque, todos los subprocesos ejecutan las mismas instrucciones y cada subproceso ejecuta la función del núcleo. La función del núcleo se basa en subprocesos para realizar el cálculo paralelo de todos los cálculos del enfoque azimutal. La Figura 11 presenta el diagrama de flujo del algoritmo BP acelerado por GPU. Después del rango compresión, la región de imagen se divide en una cuadrícula de píxeles de tamaño Nf × Nphi. Si el número de cuadrículas de píxeles en el área de imagen es inferior a 500 500 ×, entonces se asigna una cantidad de subprocesos para ser responsables de calcular los datos de eco de estos cuadrículas de píxeles y acumularlas en las posiciones correspondientes. Estas cuadrículas de píxeles se dividen en varias subregiones para el procesamiento en paralelo si los píxeles son más de 500 500 ×. Cada subproceso de la GPU es responsable de procesar un punto de un píxel, incluido el cálculo de distancia, compensación de fase, etc. Calcula la ubicación actual del píxel y acumula el eco en la cuadrícula de píxeles correspondiente. Una vez que cada bloque ha completado el cálculo de los datos de su subregión, los datos se transfieren a su posición respectiva en la imagen final. Finalmente, se genera la imagen completa.

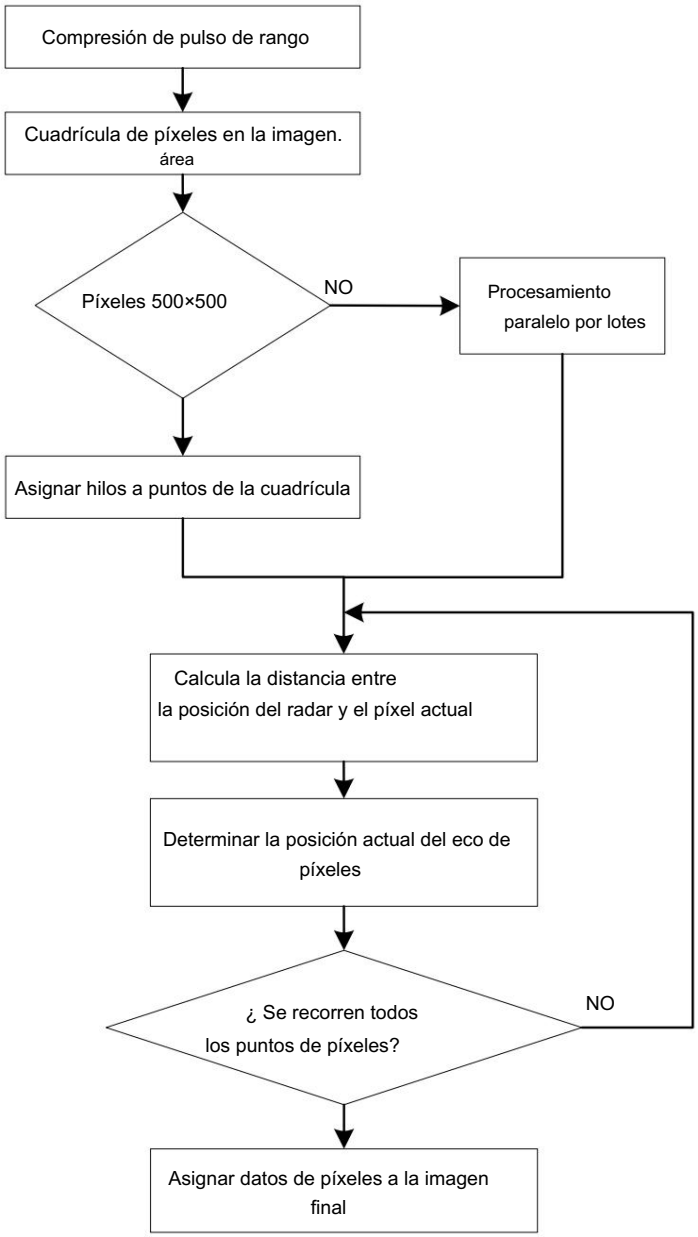


Figura 11. Diagrama de flujo del algoritmo BP acelerado por GPU.

Figura 11. Diagrama de flujo del algoritmo BP acelerado por GPU.

4. Resultados y discusión

4.1. Validación del método SBR acelerado por GPU

4.1. Validación del método SBR acelerado por GPU

La implementación del método computacional electromagnético acelerado por GPU y el método de imágenes de BP acelerado por GPU se presentó anteriormente, como se describe en este artículo. En esta sección, la validez del SBR acelerado por GPU que combina aproximaciones GO con PO se verifica en comparación con RL-GO en el software FEKO v2020. Tomando como ejemplo un caza F-22 a gran escala, el RL-GO está configurado con dos tipos de densidades de rayos, uno es λ/10 y el otro es λ/100. Se hace una comparación del RCS de campo lejano con el eléctrico /10 y el otro es /100 . Se realiza una comparación del RCS de campo lejano con la frecuencia de onda tromagnética eléctrica f0 = 3GHz. La disección del modelo produjo 2444 triángulos. Frecuencia de onda tromagnética 3GHz. La disección del modelo produjo 2444 triangulares f = mallas, y el árbol BVH se construyó para producir 4887 nodos de hojas. El modelo CAD es mallas, y el árbol BVH se construyó para producir 4887 nodos de hojas. El modelo CAD es como en la Figura 12. El tiempo de cálculo y el costo de la memoria se enumeran en la Tabla 1. Nuestra computadora como en la Figura 12. El tiempo de cálculo y el costo de la memoria se enumeran en la Tabla 1. La CPU de nuestra computadora es una Intel (R) Core (TM) i5-12490F de 12.ª generación con una velocidad de referencia de 3,00 GHz. La CPU es un Intel(R) Core (TM) i5-12490F de 12.ª generación con una velocidad de referencia de 3,00 GHz, fabricado por Intel Corporation para productos de edición especial de China. La GPU es una fabricado por Intel Corporation para productos de edición especial de China. La GPU es una NVIDIA GeForce RTX 2080 con 8,0 GB de memoria GPU dedicada, fabricada por NVIDIA GeForce RTX 2080 con 8,0 GB de memoria GPU dedicada, fabricada por NVIDIA Corporation en China continental. NVIDIA Corporation en China continental.

Machine Translated by Google
Electrónica 2024, 13, x PARA REVISIÓN POR PARES 15 de 24
Electrónica 2024, 13, 3062 15 de 24

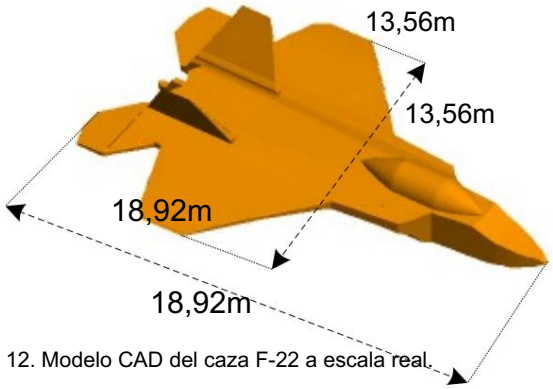


Figura 12. Modelo CAD del caza F-22 a escala real.

Figura 12. Modelo CAD del caza F-22 a escala real.
Figura 12. Modelo CAD del caza F-22 a escala real.

La Figura 13 muestra la comparación de la distribución angular RCS de un F-22 a escala real. caza obtenido mediante nuestro método SBR acelerado por GPU y el método RL-GO paralelo en la Tabla 1. Una FEKO. En la Figura 13a,b, se observa que el RCS calculado por nuestra GPU acelerada generalmente concuerda con el del método RL-GO en FEKO.

La Figura 13 muestra la comparación de la distribución angular RCS de un F-22 a escala real.

	Tiempo de cálculo (s)			Memoria (MB)		
Polariza-ción	Este papel	RL-GO	RL-GO	Este papel	RL-GO	RL-GO
	λ /10	λ /10	λ /100	λ /10	λ /10	λ /100
VV	207.96	186,4	2303.3	557.6	130,9	181,4
HH	433.195	103,1	1427.43	552,1	136,4	185.1

luchador obtenido por nuestro método SBR acelerado por GPU y el método paralelo RL-GO en este documento

Existe una ligera diferencia entre nuestro método SBR acelerado por GPU y el RL paralelo: este documento

FEKO. En la Figura 13a,b, se observa que el RCS calculado por nuestra GPU acelerada

Método GO con una densidad de rayos de λ /10

El método SBR generalmente concuerda con el método RL-GO en FEKO.

el método RL-GO con una densidad de rayos de λ también tiene en cuenta el campo de difracción [53].

Figura 13a VV Hay una

ligera diferencia entre nuestro método SBR acelerado por GPU y la cuenta para el RL- Figura 13b HH

Método GO con una densidad de rayos del λ /100 en FEKO debido al efecto de la densidad del rayo, y /100

método RL-GO con una densidad de rayos de cuenta. λ también tiene en cuenta el campo de difracción [53].

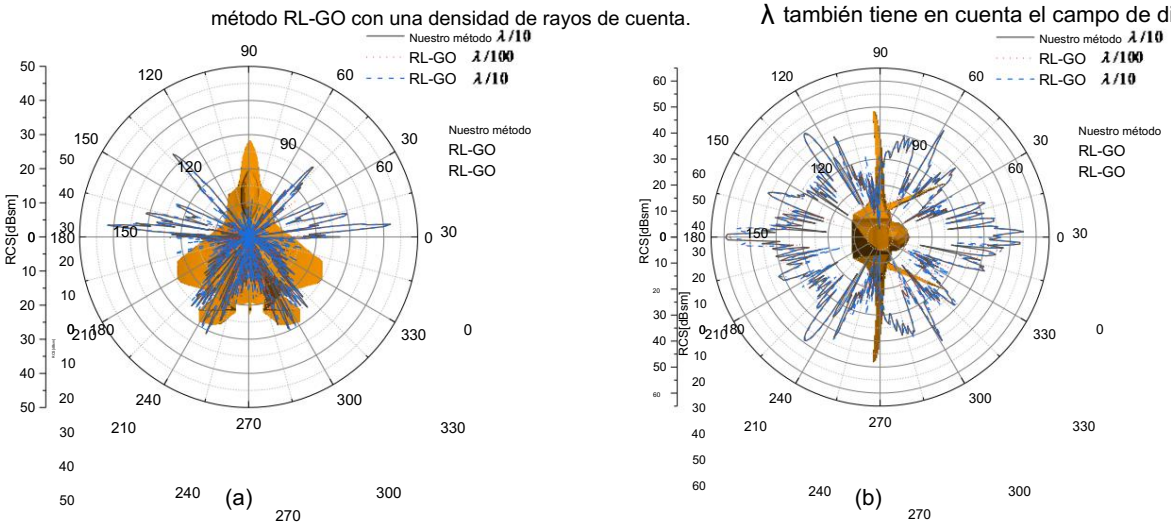


Figura 13. Distribución angular RCS de un caza F-22 a escala real: (a) ángulo de incidencia θ = 90° , ángulo de azimut φ=0 ° ~ 360 ° , polarización VV; (b) ángulo de incidencia θ = 90° , ángulo de azimut φ=0 ° ~ 360 ° , polarización HH.

Figura 13. Distribución angular RCS de un caza F-22 a escala real: (a) ángulo de incidencia θ = 90° , ángulo de azimut φ=0 ° ~ 360 ° , polarización VV; (b) ángulo de incidencia θ = 90° , ángulo de azimut φ=0 ° ~ 360 ° , polarización HH.

Tabla 1. Una comparación del tiempo de cálculo y el costo de memoria para un caza F-22 a gran escala.

Tabla 1. Una comparación del tiempo de cálculo y el costo de memoria para un caza F-22 a gran escala.

	Tiempo de cálculo (s)			Memoria (MB)		
Polariza-ción	Este papel	RL-GO	RL-GO	Este papel	RL-GO	RL-GO
	λ /10	λ /10	λ /100	λ /10	λ /10	λ /100
VV	207.96	186,4	2303.3	557.6	130,9	181,4
HH	433.195	103,1	1427.43	552,1	136,4	185.1

Método RL-GO con una densidad de rayos de cuenta. Este papel también tiene en cuenta el campo de difracción [53]

Figura 13a VV Polariza-

Figura 13b ción HH

4.2. Simulaciones de imágenes ISAR

4.2. Simulaciones de imágenes ISAR

En esta sección, los resultados de las imágenes ISAR de objetivos de aeronaves representativos se presentan y discutido. La matriz del campo retrodispersado de objetivos de aeronaves representativos es calculado mediante nuestro método SBR acelerado por GPU, y luego los ecos se enfocan para obtener imágenes ISAR utilizando el algoritmo BP acelerado por GPU desarrollado en este artículo. A modo de comparación, nuestro algoritmo BP acelerado por GPU también se aplicó para enfocar el

El método RL-GO de FEKO bajo la condición de la misma densidad de rayos, e incluso supera al RL-GO en ángulos específicos. Por ejemplo, en la Figura 16a,d, se puede ver que la diferencia en la dispersión en la parte del motor del avión es más obvia, y se puede ver que las diferencias en la dispersión en el área del motor del avión son más obvio en la Figura 16c,f; La Figura 16f tiene un fuerte desorden que abruma la información, como las características estructurales del avión, y los resultados de la Figura 16f no son tan buenos como la comparación. También se aplicó nuestro algoritmo BP acelerado por GPU para enfocar la retrodispersión. resultados de la Figura 16c. Cuando la densidad del rayo es alta, los ecos pueden registrar el campo empleando el método RL-GO en el software FEKO v2020 para obtener ISAR enfocado.

La Figura 14 muestra el modelo CAD de un modelo de avión A380 a escala con dimensiones. Este efectivo tiene 3770 triángulos. Las Figuras 15a-c ilustran tres configuraciones de observación típicas. rangos de escaneo azimutal. El ángulo de incidencia se fija en θ = 60◦ . El

Los parámetros de imágenes ISAR para las Figuras 16 y 18 se establecen como en la Tabla 2.

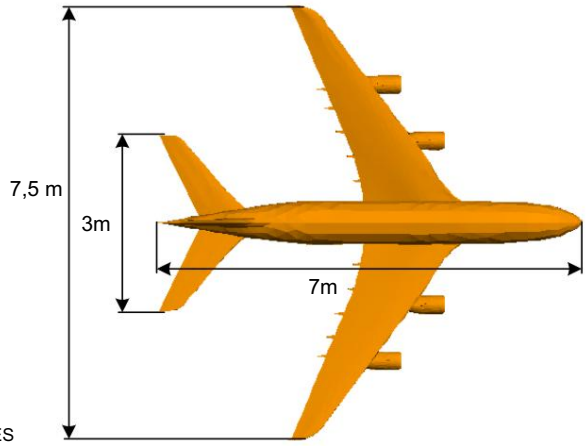
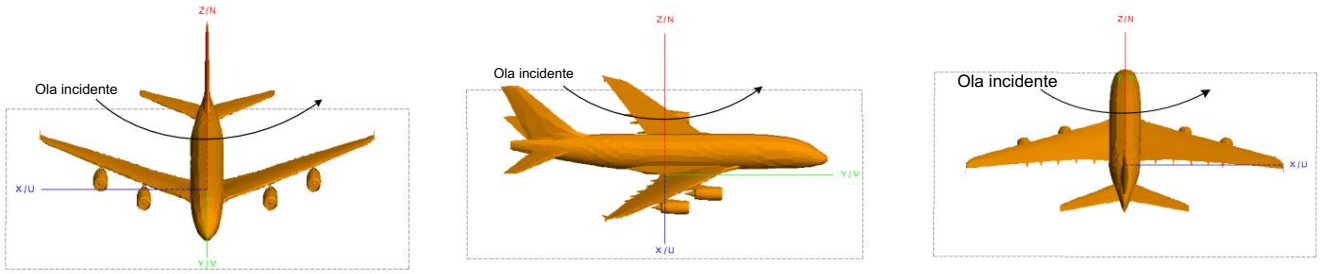


Figura 14. Modelo de diseño asistido por ordenador (CAD) y dimensiones de un modelo de avión A380 a escala.

Figura 14. Modelo de diseño asistido por ordenador (CAD) y dimensiones de un modelo de avión A380 a escala.



(a) (b) (C)

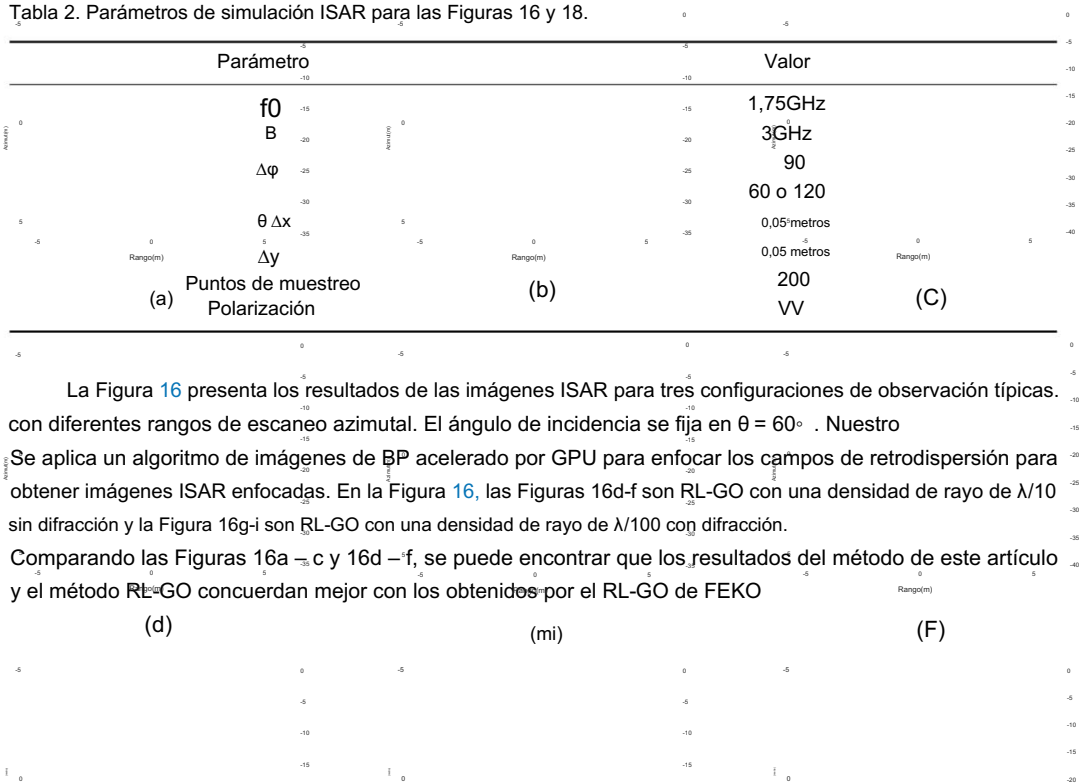
Figura 15. Tres configuraciones de observación típicas con diferentes ángulos de azimut bajo inci- Figura 15. Tres ángulo de dependencia θ = 60◦. (a) φ = 45◦ 135◦; (b) φ = −45◦ ; (c) φ = −45◦ configuraciones de observación típicas con diferentes ángulos de azimut bajo fijo ángulo de incidencia θ = 60◦. (a) φ = 45◦ 135◦; (b) φ = −45◦ ; (c) φ = −45◦ 135◦ .

Tabla 2. Parámetros de simulación ISAR para las Figuras 16 y 18.

Parámetro	Valor
f0	1,75GHz
B	3GHz
Δφ	90
θ	60 o 120
Δx	0,05 metros
Δy	0,05 metros
Puntos de muestreo	200
Polarización	VV

La Figura 16 presenta los resultados de las imágenes ISAR para tres configuraciones de observación típicas. con diferentes rangos de escaneo azimutal. El ángulo de incidencia se fija en θ = 60◦ . Nuestro Se aplica un algoritmo de imágenes de BP acelerado por GPU para enfocar los campos de retrodispersión para obtener imágenes ISAR enfocadas. En la Figura 16, las Figuras 16d-f son RL-GO con una densidad de rayo de λ/10 sin difracción y la Figura 16g-i son RL-GO con una densidad de rayo de λ/100 con difracción.

Comparando las Figuras 16a – c y 16d – f, se puede encontrar que los resultados del método de este artículo y el método RL-GO concuerdan mejor con los obtenidos por el RL-GO de FEKO



(d) (mi) (F)

método bajo la condición de la misma densidad de rayos, e incluso supera al RL-GO en ángulos específicos. Por ejemplo, en la Figura 16a,d, la diferencia en la dispersión en el motor parte del avión puede verse más obvia, y las diferencias en la dispersión en el área del motor del avión puede verse más obvia en la Figura 16c,f; Figura 16f tiene un fuerte desorden que abruma la información, como las características estructurales del avión, y los resultados de la Figura 16f no son tan buenos como los resultados de la Figura 16c. Cuando el La densidad del rayo es λ/100, los ecos pueden registrar las características estructurales geométricas de el avión en detalle, por lo que la Figura 16g-i tiene muy buenos resultados de imágenes en comparación con las Figuras 16a – cy 16d – f, lo que confirma la efectividad de las imágenes de BP aceleradas por GPU inci fijos en este artículo desde un lado. La Tabla 3 muestra el tiempo de cálculo y el pico. comparaciones de memoria para la Figura 16.

Figura 15. Tres configuraciones de observación típicas con diferentes ángulos de azimut bajo algoritmos ángulo de dependencia θ = 60 . (a) φ= 45 ~ 135 ; (b) φ= −45 ~ 45 ; (c) φ= −135 ~ 45 .

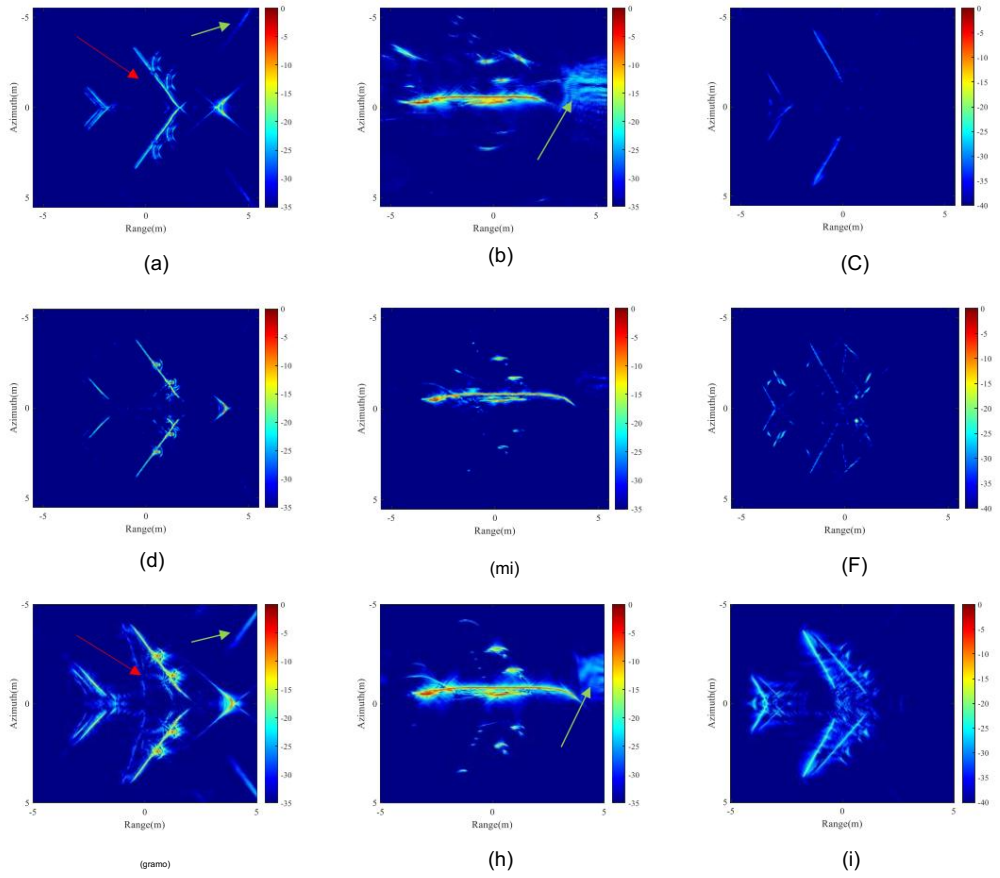


Figura 16. Resultados de imágenes ISAR utilizando el algoritmo de imágenes de BP acelerado por GPU bajo incidencia fija ángulo θ = 60 . En (a – c), los campos de retrodispersión se calculan mediante nuestro método SBR acelerado por con una densidad de rayos de λ/10; en (d – f), los campos de retrodispersión se obtienen mediante el método RL-GO de FEKO. GO con un método con una densidad de rayos de λ/10; en (g – i), los campos de retrodispersión se obtienen mediante RL-GO con un rayo densidad de rayos de λ/100. (a,d,g) son resultados para el ángulo de azimut φ= 45 ~ 135 , (b,e,h) son resultados para azimut ángulo φ= −45 ~ 45 , (c,f,i) son resultados para el ángulo de azimut φ= −135 ~ −45 .

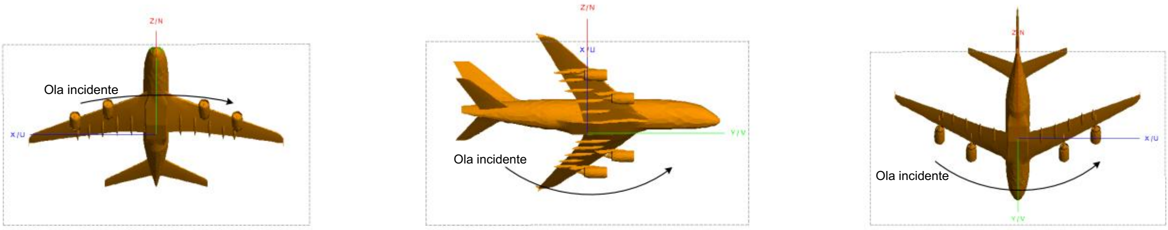
Tabla 3. Tiempo de cálculo y memoria máxima para la Figura 16.

	SBR		RL-GO	
	Tiempo (s)	Memoria (MB)	Tiempo (s)	Memoria (MB)
Figura 16a,d	2423,3 76,2	2669,8 78,4	2669,8 78,4 2623,3 72,5 2623,3 72,5	1584,2 1584,2 1614,3 1614,3 1474,1 1474,1
Figura 16b,e				177,8 177,8 179,6 179,6 179,5 179,5
Figura 16c,f				

Se puede concluir que cuando la densidad de rayos es la misma, los resultados del método en este artículo concuerdan con los de RL-GO. A continuación, analizamos la diferencia entre los resultados de los dos métodos cuando las densidades de rayos no son las mismas. En la Figura 16a, hay fuertes ecos tanto del motor como de la parte del ala unida al motor. El área indicada por la flecha roja 1 en (a) representa la posición del ala. En los ángulos de observación $\varphi = 45^\circ \sim 135^\circ$, el rayo se reflejará una vez después de golpear el ala en la posición 1. Debido a la planitud de esta parte de la estructura, el rayo posteriormente abandona el objetivo después de ser reflejado. En nuestro método SBR acelerado por GPU, la dispersión múltiple se tiene en cuenta mediante el trazado de rayos, en el que el número máximo de reflexión es 10, mientras que el campo de difracción debido a los bordes del objetivo no se tiene en cuenta, lo que conduce a una diferencia pronunciada entre la posición 1 en (a) y la posición 2 en (g).

Aparecen puntos de dispersión en las posiciones indicadas por las flechas verdes 3, 4, 5 y 6 en la Figura 16a,b,g,h. Las Figuras 16a,b son los resultados calculados mediante el método de este artículo, y la Figura 16g,h son los resultados calculados por el RL-GO con difracción con una densidad de rayo de λ/100 en FEKO. Tanto nuestro SBR acelerado por GPU como RL-GO en FEKO son métodos basados en rayos. Estos puntos de dispersión están relacionados con el mecanismo de dispersión de las ondas electromagnéticas y el mecanismo de trazado de rayos. Debido al descuido del campo de difracción resultante de los bordes del objetivo, el eco del objetivo en la Figura 16c es más débil que el de la Figura 16i. Una comparación de los resultados de las imágenes ISAR de la Figura 16a–c con las Figuras 16d–f y 16g–i demuestra la viabilidad y eficiencia de nuestro esquema combinando aceleración de GPU y el mecanismo de SBR borrado con el algoritmo BP para una simulación rápida de imágenes ISAR.

Las Figuras 17a–c ilustran tres configuraciones de observación típicas con diferentes rangos de escaneo azimutales. Resultados de imágenes ISAR para tres configuraciones de observación típicas con diferentes rangos de escaneo azimutal se muestran en la Figura 18a, b y c, respectivamente. En la Figura 18, el ángulo de incidencia se establece en $\theta = 120^\circ$, y los demás parámetros son los mismos que los de la Figura 16. En las Figuras 18a–c, los campos de retrodispersión son calculados por nuestro método SBR acelerado por GPU. En la Figura 18d–i, los campos de retrodispersión se obtienen mediante el método RL-GO correspondiente en el software FEKO v2020, donde (d–f) son RL-GO con una densidad de rayos de λ/10 y (g–i) son RL-GO con una densidad de rayos de λ/100. Dado que las densidades de rayos son las mismas y ninguno de ellos incluye el campo envolvente, los resultados para los rangos de ángulos correspondientes en (a–c) y (d–f) concuerdan bien. Una comparación de los resultados de (c) y (f) muestra que la intensidad del eco del cuerpo del avión es muy débil en (f). También hay ruido. Solo el área del motor exhibe ecos fuertes, lo que hace que el método de este artículo funcione mejor bajo las mismas densidades de rayos y puede registrar mejor la información de la estructura geométrica del objetivo. En general, las imágenes ISAR enfocadas de campos retrodispersados calculadas por nuestro método SBR acelerado por GPU concuerdan bien con los obtenidos mediante el método RL-GO de FEKO. Algunos puntos de dispersión se pueden observar, como lo indican las flechas rojas en la Figura 18a,b,g,h, como en la Figura 16. Esto se debe a la dispersión múltiple de ondas electromagnéticas. Se pueden observar desde la superficie del objetivo. Para los rayos reflejados varias veces, la historia de fase de la onda electromagnética está distorsionada. Por lo tanto, los puntos de dispersión indicados por las flechas rojas en la Figura 18a,b,g,h se pueden eliminar reduciendo los números de reflexión en el algoritmo de trazado de rayos. La Tabla 4 muestra las comparaciones de tiempo de cálculo y memoria máxima para la Figura 18.



(a) (b) (C)

Figura 17. Tres configuraciones de observación típicas con diferentes ángulos de azimut bajo ángulo de incidencia fijo $\theta = 120^\circ$. (a) $\varphi = 45^\circ \sim 135^\circ$; (b) $\varphi = -45^\circ \sim 45^\circ$; (c) $\varphi = -135^\circ \sim -45^\circ$.

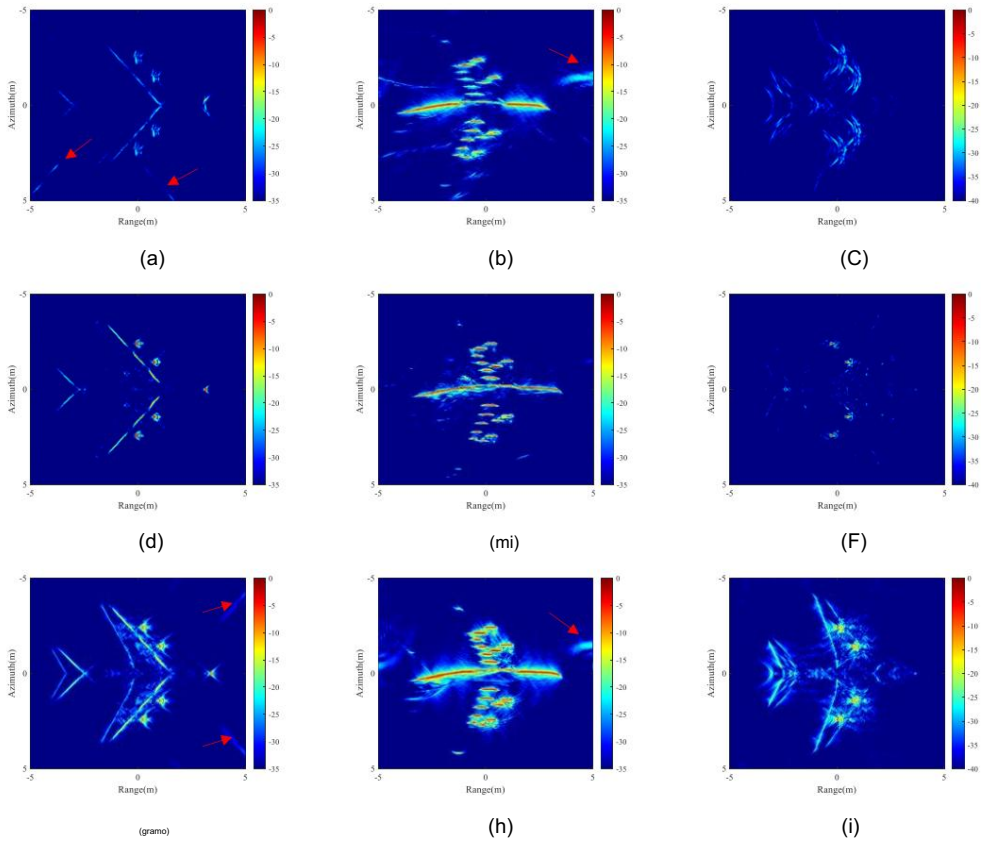


Figura 18. Similar a la Figura 16 pero con ángulo de incidencia. $\theta = 120^{o}$.

Figura 18. Similar a la Figura 16 pero con ángulo de incidencia $\theta = 120^{o}$.

Tabla 4. Tiempo de cálculo y memoria máxima para la Figura 18.

Tabla 4. Tiempo de cálculo y memoria máxima para la Figura 18.

	SBR		RL-GO	
	SBR		RL-GO	
	Tiempo (s)	Memoria (MB)	Tiempo (s)	Memoria (MB)
Figura 18a,d	2315,2	72,4	1655,6	194,6
Figura 18a,d	2647,05	74,2	1702,4	202
Figura 18b,e	2872,9	77,1	1764,2	196,9
Figura 18c,f	2872,9	77,1	1764,2	196,9
Figura 18c,f La				

Figura 19 muestra un modelo CAD de un objetivo de aeronave eléctricamente de gran tamaño, que se encuentra en la categoría de tamaño eléctricamente grande con un tamaño eléctrico de 171 × 104 longitudes de onda.

La Figura 19 muestra un modelo CAD de un objetivo de avión eléctricamente de gran tamaño, que es utilizando la categoría eléctricamente de gran tamaño con un tamaño eléctrico de 171 × 104 longitudes de onda.

Este modelo CAD tiene 712 elementos frontales. Las imágenes ISAR para este modelo se calcularon

el método en este artículo y el método RL-GO en FEKO. Las densidades de rayos para ambos Este modelo CAD

El método en este documento y el método RL-GO se establecen en (10

Las Figuras 20a a c son el método 90 0 La Figura 20 muestra tres rangos de acimut diferentes.

En las siguientes simulaciones de la Figura 21 Los parámetros de imágenes ISAR se configuran como en la

5.

En las siguientes simulaciones de la Figura 21, los parámetros de imágenes ISAR se establecen como en la Tabla 5.

La Figura 21 muestra los resultados de las imágenes ISAR de ambos métodos para imágenes eléctricamente de gran tamaño.

Tabla 5. Parámetros de simulación ISAR para la Figura 21.

objetivos de aviones, con densidades de rayos de λ/10 para este método de papel y RL-GO en FEKO.

Los Valor

parámetros que comparan las Figuras 21b y 21e, 1 y 2 en la Figura 21b son la información estructural

del avión, y 1 es el ala y 2 es la cola, mientras que la Figura 21e no puede mostrar esto

información estructural, por lo que el método de este artículo es mejor que el resultado de imágenes RL-GO

cuando φ= − 45◦ 45◦ . Comparando las Figuras 21a y 21d, el resultado de la Figura 21a es

obviamente mejor que el de la Figura 21d, donde la distorsión del historial de fase ocurre en el lugar

indicado por las flechas blancas en la Figura 21d. Información detallada sobre las alas, la nariz y

La cola del avión se puede mostrar claramente en la Figura 21a. Esto también ocurre en la Figura 21c,f,

donde apuntan las flechas blancas en la Figura 21c,f. Se puede ver claramente que las manchas dispersas

muestreo que aparecen en estos lugares indicados por las flechas no son información estructural de la aeronave.

Los puntos de

Polarización Es

un hecho bien establecido que las ondas electromagnéticas exhiben un efecto de trayectorias múltiples durante

Machine Translated by Google

Electrónica 2024, 13, x PARA REVISIÓN POR PARES 20 de 24

propagación. Este fenómeno implica que una onda incidente atraviesa una multitud de caminos para llegar a un punto de recepción designado. Cambios en los datos del historial de fases causados por estos Se pueden superponer diferentes caminos, lo que resulta en una distorsión del historial de fases. La tabla 6 muestra del tiempo de cálculo y comparaciones de memoria máxima para la Figura 21.

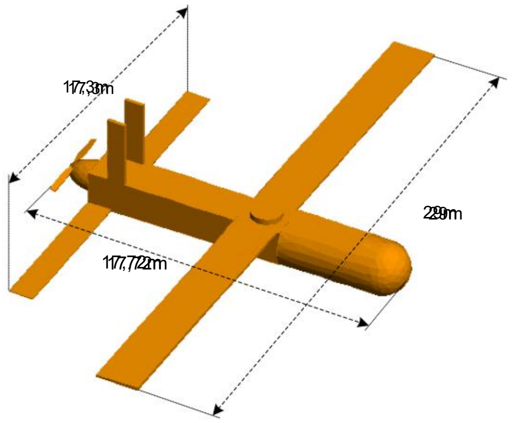
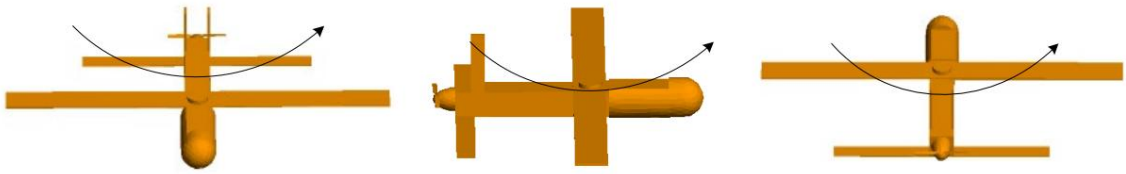


Figura 19. Modelo CAD de una aeronave eléctricamente grande.



(a) (b) (c)

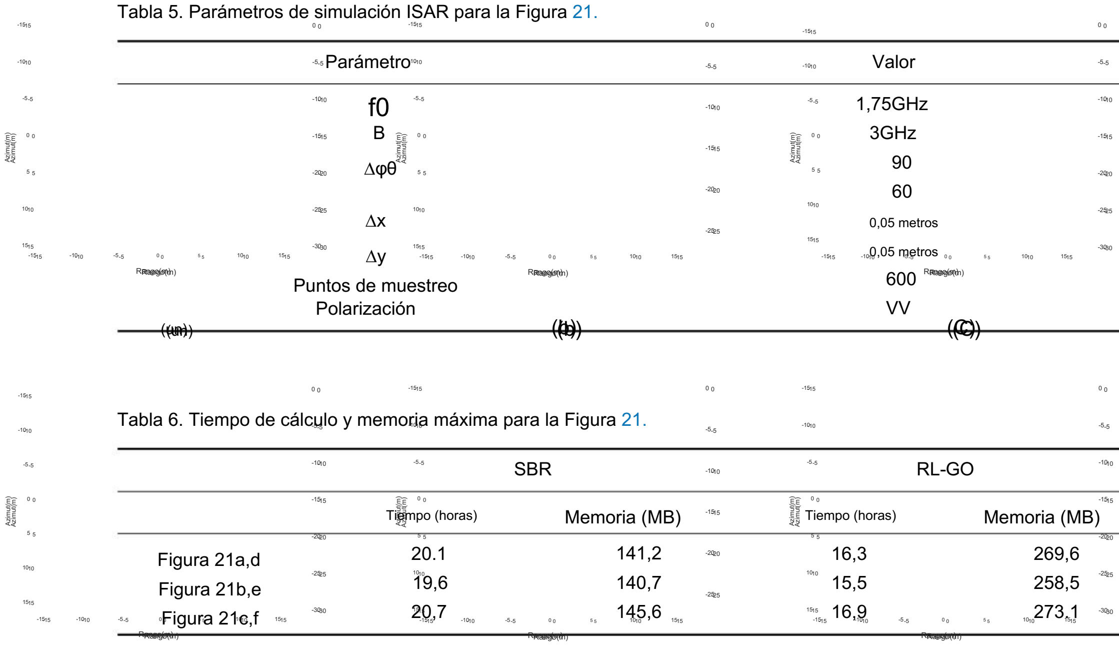
Figura 20. Tres configuraciones de observación típicas con diferentes ángulos de azimut bajo incidencia ángulo de incidencia $\theta = 60^\circ$. (a) $\varphi = 45^\circ \sim 135^\circ$; (b) $\varphi = -45^\circ \sim 45^\circ$; (c) $\varphi = -135^\circ \sim -45^\circ$.

Tabla 5. Parámetros de simulación ISAR para la Figura 21.

Parámetro	Valor
f0	1,75GHz
B	3GHz
∆φ	90
θ	60
∆x	0,05 metros
∆y	0,05 metros
Puntos de muestreo	600
Polarización	VV

Tabla 6. Tiempo de cálculo y memoria máxima para la Figura 21.

	SBR		RL-GO	
	Tiempo (horas)	Memoria (MB)	Tiempo (horas)	Memoria (MB)
Figura 21a,d	20.1	141,2	16,3	269,6
Figura 21b,e	19,6	140,7	15,5	258,5
Figura 21c,f	20,7	145,6	16,9	273,1



(a) (b) (c)

(d) (e) (f)

Figura 21. Resultados de imágenes ISAR usando el algoritmo de imágenes de BP acelerado por GPU bajo ángulos de incidencia fijos $\theta = 60^\circ$: en ángulos (a–c), los campos de retrodispersión se calculan mediante nuestro SBR acelerado por GPU; en (d–f), los campos de retrodispersión se obtienen mediante el método RL-GO de FEKO con una densidad de rayos de λ/10. (a,d) son resultados para azimut 45 ~ 135°; (b,e) son resultados para azimut −45 ~ 45°; (c,f) son resultados para azimut −135 ~ −45°. Polarización VV.

A partir de simulaciones numéricas de imágenes ISAR, se indica claramente que se pueden observar fuertes centros de dispersión, así como perfiles de objetivos, bajo un gran acimut de observación. de incidencia fijos y amplio ancho de banda. También se indica que las imágenes ISAR son muy sensibles a los ángulos de observación. Debido a la dispersión múltiple, aparecerán varios parches triangulares.

Figura 19. Modelo CAD de una aeronave eléctricamente grande.

golpeado por rayos idénticos, lo que resulta en la distorsión de la historia de fase de las ondas electromagnéticas.
La distorsión del historial de fases es un problema común con los métodos de rayos. Por tanto, los lóbulos laterales obvios
se puede observar en imágenes ISAR enfocadas. En comparación con RL-GO en FEKO v2020
software, la viabilidad y eficiencia de nuestro esquema se demuestran combinando el ángulo de decencia de GPU.
SBR acelerado con algoritmo BP para simulaciones rápidas de imágenes ISAR en gran angular
y condiciones de ancho de banda amplio.

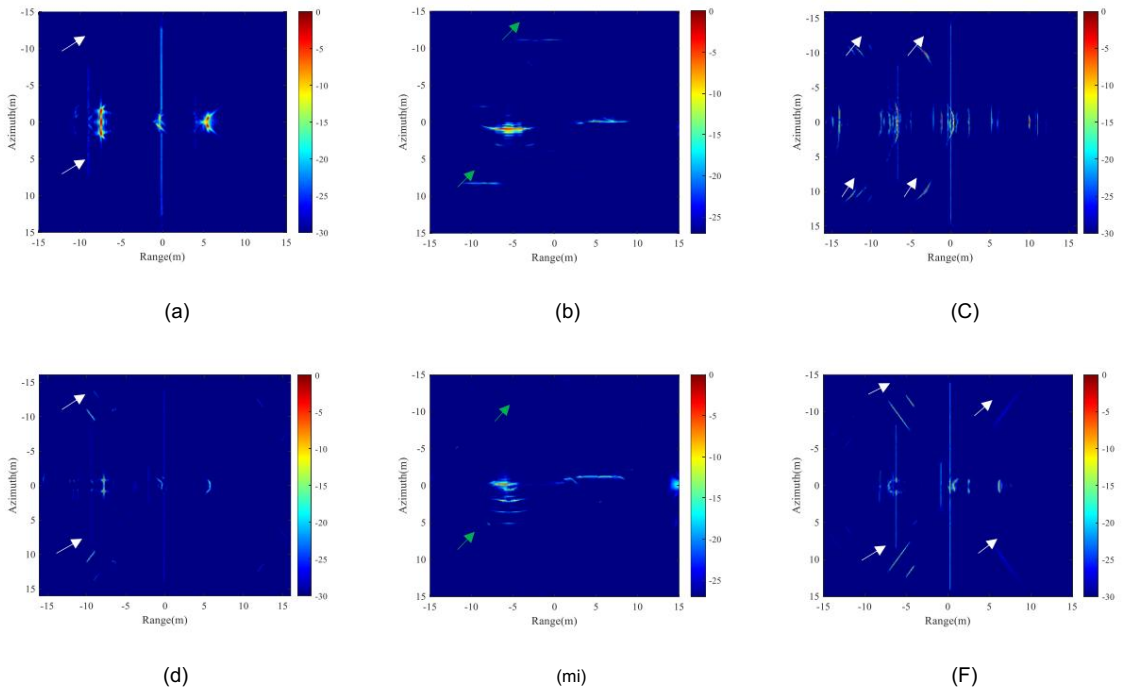


Figura 21. Resultados de imágenes ISAR usando el algoritmo de imágenes de PA acelerado por GPU bajo incidencia fija
ángulo θ = 60 ◦ en (a – c), los campos de retrodispersión se calculan mediante nuestro método SBR acelerado por GPU; en
método; en (d – f), los campos de retrodispersión se obtienen mediante el método RL-GO de FEKO con una densidad de rayos de
/10 ; (a,d) son resultados para acimut φ = 45 ◦ -135 ◦ ; (b,e) son resultados para azimut φ = −45 ◦ 45 ◦ ; (c,f) son resultados
para azimut φ = 135 ◦ -45 ◦ , Polarización VV.

5. Conclusiones

En este artículo, se presenta un novedoso método de rayos rebotantes basado en la aceleración de árboles GPU y BVH.
se presenta para simulaciones de imágenes ISAR. Emplea C++AMP para lograr el paralelo de GPU
cálculo de aceleración, en el que se construye una estructura de árbol BVH de acuerdo con el
estructura objetivo, mejorando así la eficiencia de la intersección de rayos. El método SAH se incorpora a la
división del cuadro delimitador de la escena, lo que mitiga eficazmente el impacto del cuadro delimitador.
cuadro superpuesto en la eficiencia transversal de rayos de la estructura de árbol BVH. Para eficientemente
Para realizar simulaciones de imágenes ISAR, se ha desarrollado un algoritmo de imágenes de BP acelerado basado en GPU.
desarrollado en virtud de CUDA. Se valida la precisión del SBR acelerado por GPU
comparando el RCS calculado por nuestro método SBR con el obtenido por RL-GO en
FEKO. Se demuestra que el SBR acelerado por GPU presentado muestra buena validez.
y confiabilidad. Para simulaciones de imágenes ISAR, tomando un A380 y un avión simplificado
Como ejemplo, los campos de retrodispersión se calcularon utilizando el modelo acelerado por GPU.
Algoritmo SBR bajo grandes ángulos de acimut, Δφ = 90◦, y amplios anchos de banda, 3 GHz. El
Los ecos retrodispersados se enfocan utilizando el algoritmo de imágenes de BP acelerado por GPU, y el
Las imágenes ISAR enfocadas de nuestro método SBR acelerado por GPU concuerdan bien con aquellas
del método RL-GO de FEKO, que indica la viabilidad y eficiencia de nuestra aceleración por GPU
Algoritmo de imágenes BP ISAR. Las simulaciones indican que también fuertes centros de dispersión
ya que los perfiles de los objetivos se pueden observar claramente a partir de imágenes ISAR bajo grandes ángulos de observación
y anchos de banda amplios. Se pueden observar lóbulos laterales obvios en imágenes ISAR enfocadas, debido

a la historia de fase de las ondas electromagnéticas que se distorsionan como resultado de
dispersión. Las simulaciones numéricas también indican que los resultados de las imágenes ISAR son
muy sensible a los ángulos de observación. En el futuro se ampliará el presente trabajo.
para enjambrar imágenes ISAR de objetivos, así como imágenes ISAR 3D mediante el desarrollo de un sistema eficiente
Algoritmo de simulación electromagnética mediante la combinación de SBR y PTD acelerados por GPU para
teniendo en cuenta la difracción de borde.

Contribuciones de los autores: Metodología, JC, RZ, PY y RW; software, JC; validación, JC; formal
análisis, JC; simulación de trazado de rayos, JC; investigación, JC y PY; recursos, PY; escritura—original
preparación del borrador, JC; redacción: revisión y edición, PY; visualización, JC; supervisión, PY; proyecto
administración, JC y PY Todos los autores han leído y aceptado la versión publicada del
manuscrito.

Financiamiento: este trabajo fue financiado en parte por la Fundación Nacional de Ciencias Naturales de China.
[Subvenciones Nos. 62061048, 62261054 y 62361054], en parte por Shaanxi Key Research and Development
Programa [Subvenciones Nos. 2023-YBGY-254 y 2024GXYBXM-108], en parte por el Programa Básico de Ciencias Naturales
Plan de Investigación en la Provincia China de Shaanxi (Subvención No. 2023-JC-YB-539), y en parte por el Graduado
Programa de Innovación Educativa de la Universidad de Yan'an (Subvención No. YCX2024086).

Declaración de disponibilidad de datos: Los datos presentados en este estudio están disponibles previa solicitud al
Autor correspondiente.

Agradecimientos: Los autores agradecen a los editores y revisores por sus sugerencias constructivas.

Conflictos de intereses: Los autores declaran no tener conflictos de intereses.

Abreviaturas

En este manuscrito se utilizan las siguientes abreviaturas:

PA	Retroproyección
BVH	Jerarquías de volúmenes delimitadores
C++ AMP	C++ Paralelismo masivo acelerado
CUDA	Arquitectura de dispositivo unificado de computación
CANALLA	Diseño asistido por ordenador
UPC	Unidad Central de procesamiento
electroerosión	Momento dipolar equivalente
FFT	Transformada rápida de Fourier
GPU	Unidad de procesamiento gráfico
IR	Óptica Geométrica
ISAR	Radar de apertura sintética inversa
IFFT	Transformada rápida inversa de Fourier
árbol kd	árbol K-dimensional
MOM	Método de los Momentos
[illegible]	Teoría física de la difracción
correos	Óptica Física
PEC	Conductor eléctrico perfecto
RCS	Sección de cruce de radar
RL-GO	Ray Lanzamiento de Óptica Geométrica
RMSE	Error cuadrático medio
SBR	Rayo que dispara y rebota
SAH	Heurística de área de superficie
SMS	Multiprocesadores de transmisión
SP	Procesadores de transmisión
SIMT	Instrucción única, múltiples hilos
TDSBR	Rayo de disparo y rebote en el dominio del tiempo
TDPO	Óptica física en el dominio del tiempo

Referencias

1. Bhalla, R.; Ling, H. Formación de imágenes ISAR utilizando datos biestáticos calculados a partir de la técnica de disparo y rebote de rayos. J. Electromagn. Aplicación de ondas. 1993, 7, 1271–1287. [Referencia cruzada]
2. Él, XY; Zhou, XY; Cui, TJ Simulación rápida de imágenes 3D-ISAR de objetivos en ángulos de aspecto arbitrarios mediante transformada rápida de Fourier no uniforme (NUFFT). Traducción IEEE. Propagación de antenas. 2012, 60, 2597–2602. [Referencia cruzada]
3. Zhang, K.; Wang, CF; Jin, JM Computación RCS monoestática de banda ancha e ISAR de cavidades abiertas grandes y profundas. Traducción IEEE. Propagación de antenas. 2018, 66, 4180–4193. [Referencia cruzada]
4. Prickett, M.; Chen, C. Principios del radar de apertura sintética inversa/ISAR/imágenes. En Actas de la EASCON 1980, Conferencia sobre sistemas electrónicos y aeroespaciales, Arlington, VA, EE. UU., 29 de septiembre al 1 de octubre de 1980; págs. 340–345.
5. García-Fernández, AF; Yeste-Ojeda, OA; Grajal, J. Modelo facetario de objetivos móviles para imágenes ISAR y retrodispersión de radar Simulación. Traducción IEEE. Aerosp. Electrón. Sistema. 2010, 46, 1455–1467. [Referencia cruzada]
6. Lee, JI; Yun, DJ; Kim, HJ; Yang, WY; Myung, NH Formaciones rápidas de imágenes ISAR en ángulos de múltiples aspectos utilizando el disparo y Rayos rebotantes. Antenas IEEE inalámbricas. Propaganda. Letón. 2018, 17, 1020–1023. [Referencia cruzada]
7. Guo, G.; Guo, L.; Wang, R.; Li, L. Un marco de imágenes ISAR para objetivos grandes y complejos utilizando TDSBR. Antenas IEEE Alambre. Propaganda. Letón. 2021, 20, 1928–1932. [Referencia cruzada]
8. Meng, W.; Li, J.; Xi, YJ; Guo, LX; Li, ZH; Wen, SK Un método mejorado de disparo y rebote de rayos basado en Blend-Tree para dispersión EM de múltiples objetivos en movimiento y análisis de eco. Traducción IEEE. Propagación de antenas. 2024, 72, 2723–2737. [Referencia cruzada]
9. Yang, PJ; Wu, R.; Ren, XC; Zhang, YQ; Zhao, Y. Espectros Doppler de ondas electromagnéticas dispersas desde un objeto que vuela sobre superficies marinas no lineales que varían en el tiempo. J. Electromagn. Aplicación de ondas. 2019, 33, 2175–2198. [Referencia cruzada]
10. Ling, H.; Chou, RC; Lee, SW Disparos y rayos que rebotan: cálculo del RCS de una cavidad de forma arbitraria. Traducción IEEE. Propagación de antenas. 1989, 37, 194–205. [Referencia cruzada]
11. Xu, F.; Jin, YQ Trazado de rayos analítico bidireccional para un cálculo rápido de la dispersión compuesta de un objetivo eléctrico de gran tamaño sobre un superficie aleatoriamente rugosa. Traducción IEEE. Propagación de antenas. 2009, 57, 1495–1505. [Referencia cruzada]
12. Tao, Y.; Lin, H.; Bao, H. Método de rayos de rebote y disparo basado en GPU para una predicción RCS rápida. Traducción IEEE. Propagación de antenas. 2010, 58, 494–502. [Referencia cruzada]
13. Meng, W.; Li, J.; Guo, LX; Yang, QJ Un método SBR acelerado para la predicción RCS de objetivos eléctricamente grandes. Antenas IEEE Alambre. Propaganda. Letón. 2022, 21, 1930–1934. [Referencia cruzada]
14. Baden, JM; Tripp, VK Inversión de Ray en cálculos de SBR RCS. En las actas de la 31.ª Revisión Internacional del Progreso en 2015 Electromagnético computacional aplicado (ACES), Williamsburg, VA, EE. UU., 22 a 26 de marzo de 2015; págs. 1–2.
15. Feng, TT; Guo, LX Un algoritmo de trazado de rayos mejorado para el cálculo de dispersión EM basado en SBR de objetivos eléctricamente grandes. Antenas IEEE inalámbricas. Propaganda. Letón. 2021, 20, 818–822. [Referencia cruzada]
16. Yun, KC; Fu, WC Implementación de GPU eficiente del método SBR-PO de alta frecuencia. Antenas IEEE inalámbricas. Propaganda. Letón. 2013, 12, 941–944. [Referencia cruzada]
17. Wu, R.; Wu, por; Él, PX; Guo, KY; Sheng, XQ Un algoritmo de expansión de onda plana rápida para un análisis de dispersión riguroso de Objetivos de enjambre. Traducción IEEE. Propagación de antenas. 2023, 71, 7426–7437. [Referencia cruzada]
18. Dong, CL; Guo, L.-X.; Meng, X. Un algoritmo acelerado basado en GO-PO/PTD y CWMFSM para la dispersión EM desde el barco sobre la superficie del mar y la formación de imágenes SAR. Traducción IEEE. Propagación de antenas. 2020, 68, 3934–3944. [Referencia cruzada]
19. Dong, CL; Guo, L.-X.; Meng, X.; Wang, Y. Un SBR acelerado para la dispersión EM de objetos complejos eléctricamente grandes. Antenas IEEE inalámbricas. Propaganda. Letón. 2018, 17, 2294–2298. [Referencia cruzada]
20. Meng, W.; Li, J.; Chai, SR; Xi, YJ; Wen, SK; Liu, RF Un método SBR-PTD mejorado para la dispersión EM desde un objetivo en movimiento. En actas del Simposio de la Sociedad Internacional de Electromagnético Computacional Aplicado de 2023 (ACES-China), Hangzhou, China, 15 a 18 de agosto de 2023; págs. 1–3.
21. Li, HZ; Dong, CL; Meng, X.; Guo, LX; Wei, QH Un nuevo método híbrido MoM-SBR basado en momentos dipolares equivalentes para el cálculo de dispersión EM de objetivos complejos eléctricamente grandes. En Actas de la Conferencia Internacional de 2023 sobre Tecnología de Ondas Milimétricas y Microondas (ICMMT), Qingdao, China, 14 a 17 de mayo de 2023; págs. 1–3.
22. Wang, Z.; Wei, F.; Huang, Y.; Zhang, X.; Zhang, Z. Método de imágenes RD mejorado basado en el principio de cálculo paso a paso. En Actas del 2.º Simposio Internacional SAR de China (CISS) de 2021, Shanghai, China, 3 a 5 de noviembre de 2021; págs. 1 a 5.
23. Dong, L.; Han, S.; Zhu, D.; Mao, X. Un algoritmo de formato polar modificado para SAR con misiles altamente entrecerrados. IEEE Geociencias. Sensores remotos. Lett. 2023, 20, 1–5. [Referencia cruzada]
24. Boag, AGA Imágenes de retroproyección de objetos en movimiento. Traducción IEEE. Propagación de antenas. 2021, 69, 4944–4954. [Referencia cruzada]
25. Zhang, M.; Ren, Z.; Zhang, G.; Zhang, C. Imágenes ISAR de THz utilizando un algoritmo de retroproyección con compensación de fase acelerada por GPU. J. Millim infrarrojo. Olas 2022, 41, 448–456. [Referencia cruzada]
26. Arikan, O.; Munson, DC, Jr. Un nuevo algoritmo de retroproyección para SAR e ISAR en modo foco. En Actas del Alto Speed Computing II, Los Ángeles, CA, EE. UU., 17 y 18 de enero de 1989; págs. 107-117.
27. Gong, H.; Liu, Y.; Chen, X.; Wang, C. Optimización de escena del algoritmo de retroproyección basado en GPU. J. Supercomputadora. 2023, 79, 4192–4214. [Referencia cruzada]
28. Guo, G.; Guo, L.; Wang, R.; Liu, W.; Li, L. Simulación de eco de dispersión transitoria e imágenes ISAR para un océano objetivo compuesto Escena basada en el método TDSBR. Sensores remotos 2022, 14, 1183. [CrossRef]

29. Sun, T.-P.; Cong, Z.; Él, Z.; Ding, D. Un método de óptica física iterativo acelerado en el dominio del tiempo para analizar eléctricamente Objetivos grandes y complejos. Electrónica 2022, 12, 59. [CrossRef]
30. Guo, G.; Guo, L.; Wang, R. Algoritmo de imagen ISAR que utiliza eco de dispersión en el dominio del tiempo simulado por el método TDPO. Antenas IEEE inalámbricas. Propaganda. León. 2020, 19, 1331–1335. [Referencia cruzada]
31. Zhou, J.; Han, Y. Análisis de las características de dispersión electromagnética de objetivos de plasma basándose en disparos y rayos que rebotan. método. AIP Avanzado. 2019, 9, 065106. [Referencia cruzada]
32. Gordon, WB Aproximaciones de alta frecuencia a la integral de dispersión de la óptica física. Traducción IEEE. Propagación de antenas. 1994, 42, 427–432. [Referencia cruzada]
33. Fanático, TT; Zhou, X.; Yu, WM; Zhou, XY; Cui, TJ Representaciones integrales de línea en el dominio del tiempo de campos dispersos de óptica física. Traducción IEEE. Propagación de antenas. 2017, 65, 309–318. [Referencia cruzada]
34. Liao, C. Investigación sobre modelado de dispersión de campo cercano de barcos en la superficie del mar basado en el método de alta frecuencia (en chino). Tesis de Maestría, Universidad de Ciencia y Tecnología Electrónica de China, Chengdu, China, 2021. Disponible en línea: https://kns.cnki.net/kcms2/article/abstract?v=n6BwBobH4uvU7PG733EVJdhS9-f9LApXUEAHzK60Kgv6ciotFWwf1 1njOZZqPyjAOLTnfvU-beMgAqMR8blHbC9mFOJ0F5tkD-vf1xHSqT6eY_XonBN7ouPAQRKMghtFziV-7qPBjXZhfcEiaH_ takq8r-_JoHJuM61BQbTbKyScQeIPY8_hM3AAmLBaJTfo9XIwNvOUOSI=&uni (consultado el 29 de julio de 2024).
35. Stratton, JA; Chu, L. Teoría de la difracción de ondas electromagnéticas. Física. Rev. 1939, 56, 99. [CrossRef]
36. Chu, LJ; Stratton, JA Funciones de onda elíptica y esferoidal. J. Matemáticas. Física. 1941, 20, 259–309. [Referencia cruzada]
37. Él, X.-Y.; Wang, X.-B.; Zhou, X.; Zhao, B.; Cui, T.-J. Simulación rápida de imágenes ISAR de objetivos en ángulos de aspecto arbitrarios utilizando una novedosa Método SBR. Prog. Electromagn. Res. B 2011, 28, 129-142. [Referencia cruzada]
38. Guo, G.; Guo, L.; Wang, R. El estudio sobre la dispersión de campo cercano de un objetivo bajo la irradiación de una antena mediante el método TDSBR. IEEE Acceso 2019, 7, 113476–113487. [Referencia cruzada]
39. Tang, X.; Feng, Y.; Gong, X. Algoritmo Mo M-PO/SBR basado en plataforma colaborativa y modelo mixto. Trans. Universidad de Nanjing. Aeronauta. Astronauta. 2019, 36, 589–598. [Referencia cruzada]
40. Li, J.; Meng, W.; Chai, S.; Guo, L.; Xi, Y.; Wen, S.; Li, K. Un método híbrido acelerado para la dispersión electromagnética de un Modelo compuesto de objetivo-terreno y su imagen SAR destacada. Sensores remotos 2022, 14, 6632. [CrossRef]
41. Zhu, R. Aceleración de simulaciones micromagnéticas con C++ AMP en unidades de procesamiento de gráficos. Computadora. Ciencia. Ing. 2016, 18, 53–59. [Referencia cruzada]
42. Sihai, W.; Hu, Z.; Haotian, P.; Lu, C. Paralelismo acelerado en simulación numérica con C++ AMP. En Actas del 2016 Taller: Taller sobre informática de alto rendimiento, Beijing (China), 14 a 16 de noviembre de 2016; págs. 53–55.
43. Wynters, E. Procesamiento paralelo rápido y sencillo en GPU que utilizan C++ AMP. J. Computación. Ciencia. Col. 2016, 31, 27–33.
44. Shyamala, K.; Kiran, KR; Rajeshwari, D. Diseño e implementación de multiplicación de cadenas de matrices basadas en GPU utilizando C++AMP. En Actas de la Segunda Conferencia Internacional sobre Tecnologías Eléctricas, Informáticas y de Comunicaciones (ICECCT) de 2017, Coimbatore, India, 22 y 24 de febrero de 2017; págs. 1–6.
45. Damkjær, J. Detección de colisiones BVH sin pila para simulación física; Universidad de Copenhague Universitetsparken: København, Dinamarca, 2007; Disponible en línea: http://image.diku.dk/projects/media/jesper.damkjaer.07.pdf (consultado el 29 de julio de 2024).
46. Chung, S.; Choi, M.; Tú, D.; Kim, S. Comparación de BVH y KD-tree para la aceleración GPGPU en dispositivos móviles reales. En Proceedings of the Frontier Computing: Theory, Technologies and Applications (FC 2018), Kyushu, Japón, 9 a 12 de julio de 2019; págs. 535–540.
47. Sopin, D.; Bogolepov, D.; Ulyanov, D. Construcción SAH BVH en tiempo real para escenas dinámicas de trazado de rayos. En Actas de la Гр афикон'2011, Moscú, Rusia, 26 a 30 de septiembre de 2011; págs. 74–77.
48. Huipeng, Z.; Junling, W.; Di, X.; Xiaoyang, Q. El algoritmo de retroproyección modificado para imágenes ISAR biestáticas de objetos espaciales. En Actas de la Conferencia Internacional IEEE de 2016 sobre Procesamiento de Señales, Comunicaciones y Computación (ICSPCC), Hong Kong, China, 5 a 8 de agosto de 2016; págs. 1 a 5.
49. Xiao, D. Estudio sobre imágenes ISAR de objetivos espaciales utilizando tecnología BP; Universidad de Nanjing: Nanjing, China, 2017.
50. Pu, L.; Zhang, X.; Sí.; Wei, S. Un rápido algoritmo de imágenes de retroproyección tridimensional en el dominio de la frecuencia basado en GPU. En Actas de la Conferencia de Radar IEEE de 2018 (RadarConf18), Oklahoma City, OK, EE. UU., 23 a 27 de abril de 2018; págs. 1173-1177.
51. Li, Z.; Qiu, X.; Yang, J.; Meng, D.; Huang, L.; Song, S. Un algoritmo BP eficiente basado en TSU-ICSI combinado con GPU paralela Informática. Sensores remotos 2023, 15, 5529. [CrossRef]
52. Afif, M.; Dijo, Y.; Atri, M. Aceleración de algoritmos de visión por computadora utilizando procesadores gráficos NVIDIA CUDA. Clúster. Computadora. 2020, 23, 3335–3347. [Referencia cruzada]
53. Ufimtsev, PY Ondas de borde elementales y la teoría física de la difracción. Electromagnética 1991, 11, 125-160. [Referencia cruzada]

electrónica

MDPI

Artículo

Liberación del valor empresarial: integración impulsada por la IA Toma de decisiones en los sistemas de información financiera

Alin Emanuel Artene 1,*, Aura Emanuela Domil 2 y Larisa Ivascu 1

1 Departamento de Gestión, Facultad de Gestión de la Producción y el Transporte, Universidad Politécnica de Timisoara, 14 Remus Street, 300009 Timisoara, Rumania; larisa.ivascu@upt.ro
2 Departamento de Contabilidad y Auditoría, Facultad de Economía y Administración de Empresas, Universidad Occidental de Timisoara, 300223 Timisoara, Rumania; aura.domil@e-uvt.ro
* Correspondencia: alin.artene@upt.ro

Resumen: Este artículo de investigación investiga las sinergias entre la inteligencia artificial (IA), la transformación digital (DT) y los sistemas de informes financieros dentro del contexto empresarial. El tema central explora cómo las organizaciones mejoran sus procesos de toma de decisiones mediante la integración de tecnologías de inteligencia artificial en iniciativas de transformación digital, particularmente en los informes financieros. El punto focal es comprender cómo la sinergia de estos sistemas integrados puede desbloquear un valor comercial sustancial, instigar la innovación estratégica y elevar el análisis financiero general mediante la adopción de metodologías de toma de decisiones inteligentes basadas en datos. Al aprovechar las capacidades de análisis avanzado, automatización y soporte de decisiones adaptativas, las organizaciones navegan por las complejidades de un entorno empresarial en rápida evolución , en el que las redes neuronales emergen como una herramienta valiosa para calibrar los resultados en el entorno de la contabilidad financiera, demostrando eficacia en el procesamiento de datos financieros complejos. identificar patrones y hacer predicciones, marcando el comienzo de una nueva La introducción de una matriz de beneficios de la teoría de juegos en esta herramienta de toma de decisiones de IA agrega un marco estratégico para analizar las interacciones entre los tomadores de decisiones, considerando opciones y resultados estratégicos en un contexto dinámico y competitivo.

Palabras clave: transformación digital; sistemas de toma de decisiones; Sistemas integrados; Redes neuronales; informes financieros empresariales; matriz de pagos de la teoría de juegos

Cita: Artene, AE; Domil, AE; Ivascu, L. Liberación del valor empresarial: integración de la toma de decisiones basada en IA en los sistemas de informes financieros. Electrónica 2024, 13, 3069. https://doi.org/10.3390/

electrónica13153069

Editor académico: Domenico Ursino

Recibido: 12 de julio de 2024
Revisado: 24 de julio de 2024
Aceptado: 30 de julio de 2024
Publicado: 2 de agosto de 2024

1. Introducción

La automatización, la inteligencia artificial y el análisis de datos conducen a cambios significativos en la forma en que operan las instituciones financieras. En el panorama en constante evolución de los informes financieros, las organizaciones enfrentan el imperativo urgente de aprovechar el potencial transformador de las tecnologías digitales. El impacto de la digitalización en las finanzas es innegable. El cambio de paradigma hacia la transformación digital representa un punto crítico en el que las empresas no solo deben adaptarse sino también innovar para seguir siendo competitivas en el mercado global. Un elemento central de los objetivos de esta evolución es la integración de tecnologías de inteligencia artificial (IA), que ofrecen oportunidades sin precedentes para revolucionar la gestión estratégica dentro de los sistemas de información financiera económica. El proceso de control, en este paradigma, evoluciona desde una vigilancia rígida hacia una orquestación dinámica.

Los sistemas de IA integrados equipados con algoritmos sofisticados demuestran la capacidad de aprender continuamente de flujos de datos, adaptarse a patrones en evolución y tomar decisiones en tiempo real. Esto presenta un cambio de un modelo de control determinista a un marco más adaptativo y probabilístico donde el control se ejerce a través de una gobernanza algorítmica y ciclos de retroalimentación continua.

En la era de la transformación digital pospandémica, la integración de tecnologías de inteligencia artificial es una piedra angular para las entidades económicas públicas y privadas que aspiran a navegar por el complejo panorama de la información financiera y la productividad económica con

Precisión y agilidad. La convergencia de la inteligencia artificial y la digitalización de los sistemas de información financiera [1] no sólo supone un cambio de paradigma tecnológico sino que también introduce una profunda redefinición de los mecanismos de control y los procesos de gestión. Los sistemas integrados de IA están remodelando los marcos de control y empoderando a los tomadores de decisiones en el contexto dinámico de la presentación de informes financieros en medio de la transformación digital.

A medida que las empresas se embarcan en su viaje de transformación y digitalización digital [2], la infusión estratégica y subsidiada de tecnologías de IA encierra la promesa de desbloquear un valor empresarial sin precedentes. Los sistemas tradicionales de información financiera, basados en metodologías deterministas, se encuentran ahora en una encrucijada a medida que las organizaciones buscan incorporar algoritmos avanzados de inteligencia artificial. La integración del aprendizaje automático, el análisis predictivo y la computación cognitiva en estos sistemas introduce una nueva dimensión de autonomía y adaptabilidad. A medida que la IA se convierte en un participante activo en los procesos de toma de decisiones, es necesario reevaluar los contornos de los mecanismos de control.

Los modelos jerárquicos tradicionales se ven desafiados por la naturaleza distribuida de la toma de decisiones en los sistemas integrados de IA [3], donde los algoritmos aprenden, se adaptan y contribuyen de forma autónoma . La necesidad de comprender las complejidades e implicaciones de esta integración se ha vuelto primordial, particularmente en términos de su impacto en la toma de decisiones sobre informes financieros. Los sistemas de presentación de informes tradicionales, a pesar de ser una parte integral del funcionamiento organizacional, ahora enfrentan el imperativo de evolucionar. Surge la siguiente pregunta: ¿cómo puede una integración juiciosa de las tecnologías de IA mejorar los procesos de toma de decisiones en el contexto de la presentación de informes financieros durante el proceso de transformación digital?

La Figura 1 muestra la toma de decisiones impulsada por la IA en la transformación digital. Esto incluye iniciativas de transformación digital, integración de IA, mejora de la toma de decisiones, beneficios, casos de uso, consideración, tendencias futuras, estrategias de implementación y métricas de medición . Cada una de estas direcciones es importante en la evolución de la digitalización. Evaluando cada dirección se puede afirmar lo siguiente: • Iniciativas de transformación

digital: representan pasos que contribuyen a la TD. Se pueden incluir componentes que contribuyan al éxito del proceso de TD, elementos estratégicos y oportunidades y desafíos.

• Integración de la IA: esto incluye definir cómo la IA tiene un papel importante en la TD, los diferentes tipos de tecnologías asociadas y la implementación de oportunidades y soluciones. Todo esto contribuye a un importante apoyo en el proceso de DT. • Mejora de la toma de decisiones: se

refiere al uso de datos para obtener ventajas competitivas, análisis predictivos y automatización de procesos organizacionales, y agilizar los procesos a través de una toma de decisiones efectiva.

• Beneficios: los beneficios identificados por los beneficiarios son múltiples y entre ellos se encuentran la mejora de la precisión, la eficiencia en tiempos y costos, la mejora de la experiencia del cliente y otros beneficios relacionados.

• Ejemplo: evaluando los elementos específicos en los que se puede utilizar la IA, podemos mencionar los informes financieros, la gestión de operaciones, las aplicaciones de IA de cara al cliente y

muchos otros. • Consideración: los elementos importantes que deben ser evaluados en este enfoque son Implicaciones éticas, seguridad y colaboración entre humanos y IA.

• Tendencias futuras: este campo está creciendo considerablemente y los enfoques futuros deben anticiparse a nivel organizacional a través de estrategias y enfoques bien pensados (tendencias emergentes, evolución del campo y enfoques tecnológicos). • Estrategias de implementación: los elementos de la gestión estratégica son oportunos

a nivel organizacional para un buen alineamiento con el avance tecnológico. • Métricas de medición: se debe medir todo este enfoque, y las métricas de los indicadores de desempeño y éxito para la implementación de la

IA incluyen enfoques organizacionales correctos.

Machine Translated by Google

Electrónica 2024, 13, 3069 3 de 18

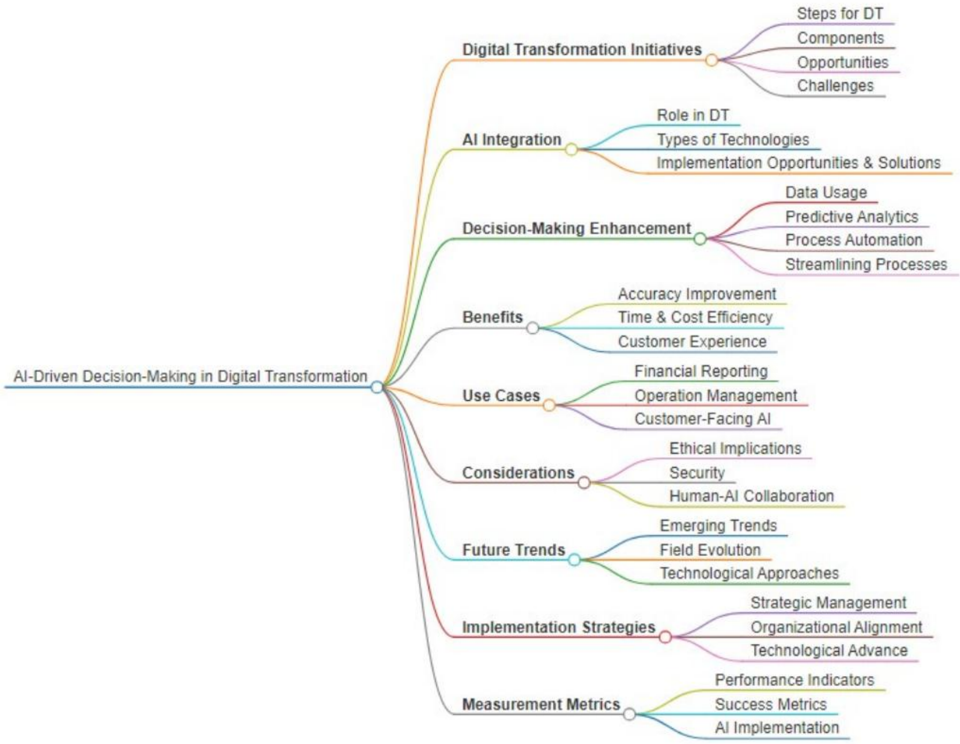


Figura 1. Mapa mental de toma de decisiones impulsado por IA.

En un panorama donde los conocimientos basados en datos impulsan el éxito organizacional, el potencial de la IA (desde la optimización de procesos hasta el descubrimiento de inteligencia procesable) invitan a las organizaciones a explorar nuevas fronteras en las capacidades de toma de decisiones. En contra de esto, las organizaciones transformador que desempeñan las tecnologías de IA. remodelación del panorama de la toma de decisiones sobre informes financieros en la era de la transformación digital. Profundizando en la complicada interacción entre la integración de la IA y los sistemas de informes financieros, buscamos proporcionar información matizada que matizados con implicaciones prácticas. Nuestro objetivo es guiar a las organizaciones cómo la IA, cuando se integra perfectamente, no solo puede de toma de decisiones en el campo dinámico de la información financiera en el mundo.

En un panorama donde los conocimientos basados en datos impulsan el éxito organizacional, los beneficios potenciales de Los beneficios de la IA, desde la optimización de procesos hasta el descubrimiento de inteligencia procesable, invitan a las deben explorar nuevas fronteras en sus capacidades de toma de decisiones. En este contexto , este artículo explora el papel En el terreno, este artículo emprende una exploración del papel transformador que la tecnología de IA está desempeñando en la Las nuevas tecnologías están desempeñando un papel en la remodelación del panorama de la toma de decisiones en materia de de la transformación digital. Al profundizar en la complicada interacción entre la integración de la IA y la naturaleza cambiante de y la naturaleza cambiante de los sistemas de información financiera, buscamos proporcionar avances teóricos interconectados y miras que conectan los avances teóricos con implicaciones prácticas. Nuestro objetivo es guiar hacia una comprensión integral de organizaciones a una comprensión integral de cómo la IA, cuando se integra perfectamente, satisface pero supera las necesidades no sólo puede satisfacer sino superar las necesidades de toma de decisiones en el campo dinámico de la era redigital financiera. portabilidad en la era digital.

2. Revisión de la literatura

2. Revisión de la literatura

La integración de la inteligencia artificial (IA) en los sistemas de información financiera ha ganado atención significativa en la literatura reciente [4] debido a su potencial para transformar los sistemas tradicionales. procesos financieros y desbloquear el valor empresarial. Los autores de [5] investigan el impacto en los sistemas de información contable para mejorar las empresas manufactureras, revelando importantes resultados positivos sistemas de información contable como en los aspectos no financieros en general. estudios han enfatizado el papel de la IA en la mejora de las finanzas toma de decisiones. Árbitro. [1] demostró la eficacia del aprendizaje automático financieras, optimizar la asignación de recursos y proporcionar planificación estratégica. La integración de redes neuronales, como se discute en [6], se ha mostrado prometedora en la automatización de tareas, mejorando la precisión de los pronósticos y a los tomadores de decisiones en el ámbito financiero.

La integración de la inteligencia artificial (IA) en los sistemas de información financiera ha recibido una ha ganado una atención significativa en la literatura reciente [4] debido a su potencial para transformar los procesos financieros tradicionales y liberar valor empresarial. Los autores de [5] investigan la integración de inteligencia artificial el impacto de la integración de la inteligencia artificial en los sistemas de información contable para el desempeño no financiero en mejorar el desempeño no financiero en las empresas manufactureras, revelando efectos significativos tanto en la eficiencia de los efectos positivos tanto en la eficiencia de los sistemas de información contable como en el incumplimiento general . Numerosos rendimiento financiero. Numerosos estudios han enfatizado el papel de la IA en la mejora de los procesos de procesos de toma de decisiones financieras. Árbitro. [1] demostró la eficacia de los algoritmos automáticos para predecir tendencias algoritmos de aprendizaje para predecir tendencias financieras, optimizar la asignación de recursos e información valiosa para la proporcionando información valiosa para la planificación estratégica. La integración de redes neuronales, como discutido por [6], se ha mostrado prometedor en la automatización de tareas, mejorando la precisión de los pronósticos y apoyando y apoyar a los responsables de la toma de decisiones en el ámbito financiero.

El papel del big data y el aprovechamiento de la toma de decisiones impulsada por la IA y la transformación digital El papel de informes financieros libera valor empresarial, alineándose con las preferencias de las partes interesadas en los sistemas de información financiera desbloquean el valor empresarial [5] se alinean con las partes interesadas y abordan los desafíos y oportunidades en el panorama financiero en evolución [5]. el impacto transformador del big data se extiende de las empresas. En el contexto de la economía moderna, el impacto transformador del big data

del big data y el aprovechamiento de la toma de decisiones impulsada por la IA y la transformación digital en los sistemas preferencias y abordar los desafíos y oportunidades en el cambiante mundo financiero. En el contexto de la economía moderna, desde [5]. En el contexto de la economía moderna, el impacto transformador del big data va más allá de su uso común por parte

cómo la utilización de los datos por parte de otros influye en los resultados del mercado [5]. Esta exploración implica incorporar big data a las teorías económicas y financieras contemporáneas. En particular, una aplicación implica aprovechar los macrodatos para mejorar la toma de decisiones de los participantes en los mercados financieros, influyendo en los precios de las empresas, el costo del capital y la dinámica de inversión.

El concepto de transformación digital en la información financiera se ha explorado ampliamente en la literatura. Académicos como los autores de [7,8] discutieron la evolución de los sistemas de información financiera desde procesos manuales hasta plataformas digitales avanzadas. El cambio hacia soluciones basadas en la nube, la automatización de las tareas de entrada de datos y el uso de herramientas de análisis avanzadas se han identificado como componentes clave de la transformación digital en la presentación de informes financieros [9].

La investigación de Sorensen [10] investiga la importancia de alinear las decisiones financieras con las preferencias de las partes interesadas. Árbitro. [10] destaca la necesidad de herramientas que consideren las perspectivas de diversas partes interesadas, como la dirección de la empresa y los accionistas.
La integración de herramientas de toma de decisiones impulsadas por IA, informadas por redes neuronales, ofrece un mecanismo para alinear las opciones estratégicas con las preferencias de las partes interesadas clave.

La literatura también reconoce los desafíos asociados con la integración de la IA y la transformación digital en la información financiera. Se han debatido las preocupaciones relacionadas con la privacidad de los datos, el cumplimiento normativo y la necesidad de profesionales capacitados capaces de navegar en la intersección de las finanzas y la IA [3,11]. Sin embargo, los estudios también destacan las oportunidades de ahorro de costos, mejoras de eficiencia y ventajas estratégicas que surgen de una implementación exitosa [12]. Numerosos estudios enfatizan la importancia de los KPI financieros en la medición del desempeño y la toma de decisiones estratégicas dentro de las organizaciones [13]
Métricas como el retorno de la inversión (ROI), las ganancias por acción (EPS) y el margen de beneficio se reconocen como indicadores críticos de la salud financiera. Los investigadores sostienen que la incorporación de estos KPI en redes neuronales puede mejorar la precisión de las predicciones financieras y los sistemas de apoyo a las decisiones [14,15].

Al examinar la intersección de la teoría de juegos y la toma de decisiones financieras [16] se explora cómo las interacciones estratégicas entre los participantes del mercado, influenciadas por la asimetría de la información y la competencia, impactan los resultados financieros. Es posible que el estudio no se centre directamente en los KPI, pero proporciona información sobre la dinámica de las decisiones. Si bien puede que no exista una extensa literatura que combine explícitamente los KPI financieros, la toma de decisiones y la teoría de juegos, algunos estudios proporcionan una base para comprender la interconexión de estos conceptos. Abordar la transparencia y la interpretabilidad de los modelos de decisión de IA, Ref. [17] analiza la importancia de la explicabilidad para ganarse la confianza y la aceptación de las herramientas de IA en la toma de decisiones, especialmente en ámbitos sensibles como la atención sanitaria y las finanzas.

La revisión de la literatura indica un creciente cuerpo de investigación que enfatiza las diversas aplicaciones, desafíos y consideraciones éticas asociadas con la integración de herramientas de IA en la toma de decisiones. Comprender el impacto de la IA en diversos dominios proporciona una base para futuros desarrollos en la creación de un sistema de apoyo a las decisiones más eficaz, transparente y ético .

En los últimos años, las empresas de contabilidad y auditoría han aprovechado el potencial de la IA para revolucionar los procesos tradicionales dentro de las instituciones financieras y han desarrollado aplicaciones de aprendizaje automático en las finanzas. JP Morgan Chase, líder mundial en servicios financieros, desarrolló la plataforma Contract Intelligence, o COiN, una herramienta impulsada por IA diseñada para revisar documentos legales y extraer cláusulas y puntos de datos críticos, reduciendo hasta 360.000 h al año. consumiendo importantes recursos humanos y mejorando la precisión y escalabilidad de las operaciones, estableciendo un nuevo estándar de eficiencia en la información financiera. Otro ejemplo convincente proviene de Deloitte, una potencia en servicios de auditoría y aseguramiento que empleó ACL Analytics, un sofisticado software de análisis de datos, para mejorar sus procesos de auditoría. Esta herramienta aprovecha el poder de la inteligencia artificial y el aprendizaje automático para examinar grandes volúmenes de datos financieros, identificando anomalías, tendencias y patrones que justifican un mayor escrutinio. Al integrar redes neuronales, los auditores pueden concentrar sus esfuerzos en áreas de alto riesgo, mejorando así la calidad y profundidad de sus auditorías y

Brindar a los clientes recomendaciones valiosas basadas en datos, fomentando una toma de decisiones y una planificación estratégica más informadas.

3. Materiales y métodos

Este trabajo de investigación utiliza una metodología de investigación fundamental integral para desbloquear el valor empresarial mediante la integración de la toma de decisiones basada en IA con la transformación digital en los sistemas de informes financieros. El proceso lógico descrito en esta metodología sirve como hoja de ruta paso a paso, guiando la investigación sobre la intersección entre la inteligencia artificial (IA) y la transformación digital en la economía en el contexto de la digitalización de los informes financieros. El objetivo principal es desarrollar un marco teórico que mejore los procesos de toma de decisiones mediante la asimilación de tecnologías avanzadas de inteligencia artificial en los sistemas de información financiera existentes. A lo largo de esta investigación pasamos por etapas sistemáticas, comenzando con la identificación del problema de investigación, seguida de una extensa revisión de la literatura para recopilar información relevante sobre inteligencia artificial , transformación digital y sistemas actuales de información financiera. Luego se trazó el fundamento teórico sintetizando el conocimiento existente e identificando lagunas en la comprensión actual. Posteriormente, los autores formularon un diseño de investigación para guiar la investigación empírica.

Integrar la toma de decisiones basada en IA con la transformación digital en los sistemas de información financiera requiere una cuidadosa consideración de las dimensiones tecnológicas, organizativas y éticas . La metodología aborda estas cuestiones incorporando un enfoque multidisciplinario, garantizando una comprensión holística de los desafíos y oportunidades asociados con la implementación de la inteligencia artificial en los informes financieros.

Al navegar a través de este proceso metódico, la investigación tiene como objetivo proporcionar información valiosa tanto al mundo académico como a las empresas y la industria. El resultado esperado es un marco teórico sólido que no solo aclara las sinergias entre la inteligencia artificial y la transformación digital, sino que también proporciona orientación práctica para las organizaciones que buscan mejorar sus capacidades de toma de decisiones en materia de informes financieros.

Esta investigación tiene como objetivo cerrar la brecha entre los avances teóricos y las implicaciones prácticas, desbloqueando así el valor empresarial inexplorado inherente a la integración de la toma de decisiones impulsada por la IA con la transformación digital en los sistemas de informes financieros.

4. Herramientas de transformación digital para el proceso de toma de decisiones

Las herramientas de transformación digital para la presentación de informes financieros pueden respaldar significativamente el proceso de toma de decisiones dentro de las organizaciones, desarrollando estrategias que sean resilientes a diferentes condiciones económicas. Desempeñan un papel crucial en el respaldo de los procesos de toma de decisiones en los informes financieros al proporcionar datos en tiempo real, análisis avanzados, automatización y funciones de colaboración. Estas herramientas contribuyen a una toma de decisiones más informada, estratégica y basada en datos dentro de las organizaciones. La integración de tecnologías avanzadas y herramientas digitales en los sistemas de información financiera ofrece varios beneficios que mejoran las capacidades de toma de decisiones. Los tomadores de decisiones pueden acceder a información actualizada sobre métricas financieras clave, indicadores de desempeño y tendencias del mercado, lo que permite una toma de decisiones más informada y oportuna [17,18].

Las herramientas avanzadas de análisis y visualización de datos ayudan a transformar datos financieros complejos en representaciones visuales fácilmente comprensibles. Esto ayuda a los tomadores de decisiones a identificar patrones, tendencias y anomalías, facilitando la toma de decisiones basada en datos. Las herramientas digitales a menudo incorporan capacidades de pronóstico y análisis predictivo. Al analizar datos históricos e identificar patrones, estas herramientas pueden ayudar a los tomadores de decisiones a hacer predicciones más precisas sobre tendencias y resultados financieros futuros. La incorporación de algoritmos de inteligencia artificial y aprendizaje automático (ML) mejora las capacidades de los sistemas de informes financieros . Estas tecnologías pueden proporcionar información inteligente, identificar oportunidades y respaldar la toma de decisiones mediante el análisis de grandes conjuntos de datos y facilitando la colaboración y la comunicación entre las diferentes partes interesadas involucradas en el proceso de toma de decisiones [14,15].

colaboración y comunicación entre las diferentes partes interesadas involucradas en el proceso de toma de decisiones [14,15].

Los tomadores de decisiones pueden evaluar las implicaciones financieras de diversas decisiones, asegurando la alineación con los requisitos regulatorios y las estrategias de mitigación de riesgos y permitiendo que las nario planificación al permitir a los tomadores de decisiones modelar diferentes escenarios financieros y alinearlos con los de interacciones estratégicas entre la planificación de múltiples decisiones al permitir a los tomadores de decisiones modelar Los creadores, a menudo en situaciones competitivas, pueden ser empleados para apoyar a los gerentes y evaluar su el proceso de toma de decisiones. En contextos de negocios, se puede utilizar para modelar y los creadores, a menudo en situaciones competitivas, a múltiples partes interesadas, como competidores y titulares en el proceso de toma de decisiones. En contextos empresariales, se puede proveedores y clientes [19,20]. Analizar de decisiones que involucren a múltiples partes interesadas, como competidores,

La teoría de juegos y el aprendizaje automático brindan conocimientos estratégicos para la evaluación de riesgos al analizar diversos conjuntos de datos e identificar tendencias potenciales para ayudar a la teoría de juegos y el Los directivos y accionistas anticipar las acciones y reacciones de los demás [18]. Ellos y ayudando ayuda en la toma de decisiones que consideran las respuestas probables de los competidores y colaboración, la toma de laboratorios, lo que lleva a estrategias más informadas. Redes neuronales, un subconjunto de máquinas que conducen a El aprendizaje y la inteligencia artificial son expertos en el reconocimiento y la predicción de patrones. En la toma de Las redes rurales pueden analizar grandes conjuntos de datos, identificar patrones y hacer predicciones basadas en datos Información histórica y en tiempo real. Esta capacidad es valiosa para los gerentes y para compartir información en tiempo tenedores al evaluar diversas opciones de decisión y sus riesgos asociados. evaluar diversas opciones de

Este mapa conceptual (Figura 2), proporciona una representación visual de la interconexión. Este mapa conceptos conectados, que ilustran cómo la teoría de juegos y las redes neuronales, cuando se integran en conceptos, que sistemas de apoyo a las decisiones, contribuir a los procesos de toma de decisiones estratégicas y mejorar los sistemas de calidad general de las decisiones para las partes interesadas [21]. La combinación de teoría de juegos y calidad Las redes pueden ser particularmente poderosas. Las redes neuronales pueden analizar datos para informar decisiones tomadores de decisiones sobre los resultados probables de diferentes estrategias, mientras que la teoría de juegos puede ayudar sobre los modelar las interacciones estratégicas y la dinámica competitiva entre las partes interesadas. Por palanca: interacciones redes neuronales envejecidas para analizar datos históricos y predecir resultados, redes de tomadores de decisiones para pueden utilizar la teoría de juegos para optimizar sus estrategias basándose en una comprensión más profunda de la teoría panorama competitivo. La IA, incluidas la teoría de juegos y las redes neuronales, puede integrarse en el panorama. La IA, brindan información útil a los gerentes y sistemas de apoyo a las acciones que brindan información útil a los gerentes y titulares.

requisitos regulatorios y las estrategias de mitigación de riesgos y permitir a los sce- evaluar su impacto potencial. Análisis diferentes escenarios financieros y impacto potencial. Análisis de interacciones estratégicas entre múltiples tomadores de decisiones en se pueden emplear para apoyar a los gerentes y analizar escenarios de toma de decisiones que involucran utilizar para modelar y escenarios de toma de decisiones que involucren a múltiples partes interesadas, como competidores, proveedores y clientes [19,20].

aprendizaje automático brindan conocimientos estratégicos para la evaluación de riesgos al analizar diversos conjuntos de datos e identificar tendencias potenciales [18]. Los directivos y accionistas anticipan las acciones y reacciones de los demás [18]. Ellos ayudan en decisiones que consideren las respuestas probables de los competidores y colaboradores, estrategias más informadas. Redes neuronales, un subconjunto del aprendizaje automático, decisiones, los neu- AI son expertos en el reconocimiento y la predicción de patrones. En la toma de decisiones, las redes neuronales históricos y real. Esta capacidad es valiosa para los gerentes y accionistas cuando decisión y sus riesgos asociados.

conceptual (Figura 2), proporciona una representación visual de la interconexión ilustran cómo la teoría de juegos y las redes neuronales, cuando se integran en la toma de decisiones apoyo, contribuir a los procesos de toma de decisiones estratégicas y mejorar en general de decisión neuronal para las partes interesadas [21]. La combinación de teoría de juegos y redes neuronales. pueden ser particularmente poderosas. Las redes neuronales pueden analizar datos para informar a los tomadores de decisiones resultados probables de diferentes estrategias, mientras que la teoría de juegos puede ayudar a modelar los resultados probables de diferentes estrategias. estratégicas y dinámicas competitivas entre las partes interesadas. Aprovechando los nervios analizar datos históricos y predecir resultados, los tomadores de decisiones pueden usar juegos para optimizar sus estrategias basándose en una comprensión más profunda de la competitividad incluida la teoría de juegos y las redes neuronales, se puede integrar en sistemas de apoyo a la toma de decisiones que accionistas.

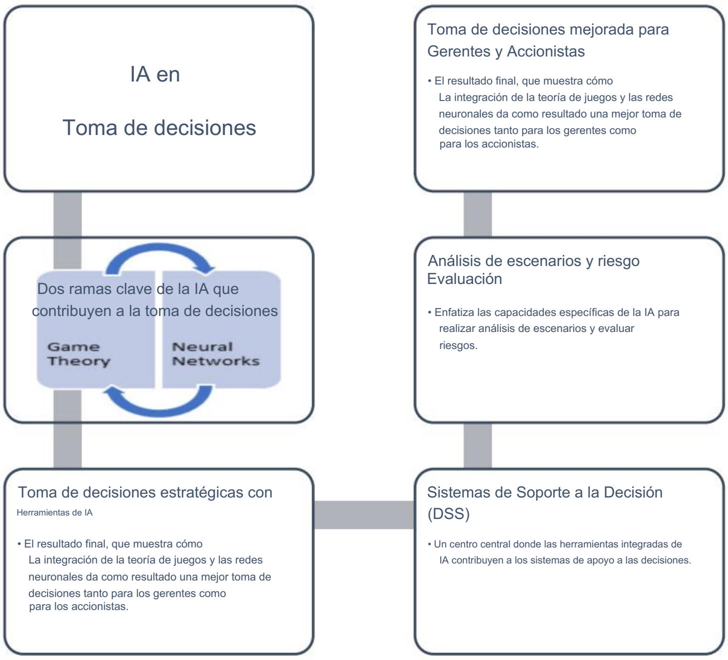


Figura 2. Mapa conceptual de la IA integrada en los sistemas de apoyo a la decisión.

Figura 2. Mapa conceptual de la IA integrada en los sistemas de apoyo a la decisión.

Estos sistemas aprovechan el análisis avanzado para mejorar los procesos de toma de decisiones.

Los gerentes pueden utilizar herramientas de inteligencia artificial en el proceso de toma de decisiones para realizar análisis de escenarios,

evaluar los resultados potenciales de diferentes decisiones en diversas condiciones. Esto ayuda a tomar decisiones más sólidas y adaptables a diferentes circunstancias.

4.1. Desarrollo de una red neuronal para integrar la IA en los procesos de toma de decisiones

El desarrollo de una red neuronal para integrar la IA en los informes financieros implica diseñar un modelo que pueda analizar e interpretar datos financieros de los balances. La elección de una red neuronal de tres capas para tareas de contabilidad y presentación de informes financieros no es una regla inherente, sino más bien una arquitectura de uso común que ha demostrado eficacia en diversas aplicaciones. Una arquitectura de tres capas es relativamente simple y más fácil de interpretar cuando la comparamos con contrapartes más profundas. Esta arquitectura tiene menos parámetros y capas, lo que aparentemente la hace más manejable. Sin embargo, esto no debería ocultar la realidad de que incluso una red neuronal de tres capas puede funcionar como una caja negra compleja. Las transformaciones no lineales inherentes y las interacciones entre capas hacen que sea difícil descifrar cómo se traducen entradas específicas en salidas. En los informes financieros, donde la transparencia y la rendición de cuentas son primordiales, la naturaleza opaca de las redes neuronales, independientemente de su profundidad, puede generar importantes desafíos de interpretabilidad. Debido a que la interpretabilidad es a menudo crucial, y como las partes interesadas necesitan comprender y confiar en los resultados del modelo, esta complejidad no es sólo una preocupación teórica sino práctica que impacta la confianza y la confiabilidad en las aplicaciones del mundo real.

Las redes neuronales de tres capas son notablemente poderosas debido a su capacidad para aproximarse a una amplia gama de funciones. Esta capacidad tiene sus raíces en el teorema de superposición de Kolmogorov, que afirma que una red neuronal de tres capas puede representar cualquier función multivariada, ya sea continua o discontinua. Esto hace que estas redes sean versátiles para diversas aplicaciones, incluidos los sistemas de informes financieros, donde es crucial capturar relaciones complejas y no lineales dentro de los datos [22,23]. Una de las principales ventajas de una red neuronal de tres capas es su relativa simplicidad e interpretabilidad en comparación con redes más profundas. Si bien puede resultar complicado interpretar el proceso de toma de decisiones de redes neuronales muy profundas, los modelos de tres capas logran un equilibrio al ser lo suficientemente complejos como para capturar patrones intrincados y, al mismo tiempo, seguir siendo más simples de analizar y depurar. Esta interpretabilidad es particularmente importante en contextos financieros donde la transparencia y la rendición de cuentas son primordiales [24,25].

Los modelos más simples también son menos propensos al sobreajuste y pueden ser más sólidos, especialmente cuando se trata de datos limitados. Para muchas tareas de contabilidad y presentación de informes financieros, un número moderado de capas ocultas puede capturar eficazmente los patrones y relaciones subyacentes en los datos.

Diseñamos el modelo matemático para una red neuronal con 31 unidades de entrada, 16 unidades ocultas y 2 unidades de salida de la siguiente manera:

x1, x2, ..., x31 para ser las características de entrada

h1, h2, ..., h16 para ser las unidades de capa ocultas

y1, y2 serán las unidades de la capa de salida

El modelo matemático de la red neuronal se puede expresar de la siguiente manera:

Cálculo de capa oculta:

$$hj = \text{ReLU} \sum i = 131wij(1)xi + bj(1) \text{ , para } j = 1, 2, \ldots, \text{dieciséis}$$

donde ReLU(a) = max(0, a) es la función de activación de la Unidad Lineal Rectificada.

Cálculo de la capa de salida:

$$yk = \text{Softmax} \sum j = 116wjk(2)hj + bk(2) \text{ , para } k = 1, 2$$

eran

$$\text{Softmax}(a)i = \frac{eai}{\sum j = \text{es la función de activación de SoftMax. } \sum 1 \text{ unidad}}$$

$$\text{Softmax}(\ ) = \sum = \frac{}{\sum}$$ es la función de activación de SoftMax.

Los parámetros del modelo (pesos y sesgos) se aprenden durante el entrenamiento Los parámetros del modelo (pesos y sesgos) se aprenden durante el entrenamiento
proceso para minimizar una función de pérdida específica, típicamente asociada con la tarea en cuestión proceso
para minimizar una función de pérdida específica, típicamente asociada con la tarea en cuestión (p. ej.,
regresión). El entrenamiento implica ajustar los pesos y sesgos de clasificación o
algoritmos de optimización como el descenso de gradiente estocástico (SGD).

El modelo se entrena realizando iterativamente propagación hacia adelante, calculando El modelo se
pérdida y luego usar la propagación hacia atrás para actualizar los parámetros del modelo. El entrenamiento
El proceso continúa hasta que el modelo converge o hasta un número predeterminado de épocas. El proceso
logrado. Una vez entrenada, la red neuronal se puede utilizar para realizar predicciones sobre nuevos datos .
propagación hacia adelante con los parámetros aprendidos. Esta aplicación matemática realiza la
El enfoque forma la base de la capacitación utilizando una red neuronal de tres capas para informes financieros.
y, en última instancia, predecir los KPI financieros en la capa de producción. Se pueden realizar ajustes en la
en función de las características específicas de sus datos y requisitos de la tarea. realizado

En entornos financieros donde los datos pueden ser limitados y los recursos computacionales pueden ser
restringida, una arquitectura de tres capas logra un equilibrio entre complejidad y eficiencia.
explota. Las redes extremadamente profundas pueden sufrir problemas de gradiente que desaparecen o
problemas durante el entrenamiento, lo que dificulta que el modelo aprenda de manera efectiva. Un tres durante
Las capas es menos propensa a estos problemas, especialmente cuando se usan funciones de activación
ReLU (unidad lineal rectificada) que mitigan el problema del gradiente de fuga [26].

La capa de entrada debe adaptarse para capturar información relevante de la balanza. La capa de
hoja. Como se ve en la Figura 3. se explica con más detalle en el Apéndice A. Cada nodo en la capa de
entrada corresponde a una característica o variable específica del balance general, como activos,
pasivos, capital, impuestos, depreciación, ingresos netos, etc. La inclusión de características relacionadas con la
estructura del balance permite que la red neuronal aprenda patrones y relaciones dentro de los datos
financieros. Esta información es crucial para comprender la salud financiera y la posición de
la entidad económica. Aunque se trata de una propuesta de modelo para una economía
efectividad de nuestra red neuronal depende de la calidad y relevancia de la
la contabilidad proporcionada. Se debe garantizar que las características seleccionadas en la capa de entrada
sean significativos y que haya datos suficientes para entrenar el modelo de manera efectiva.

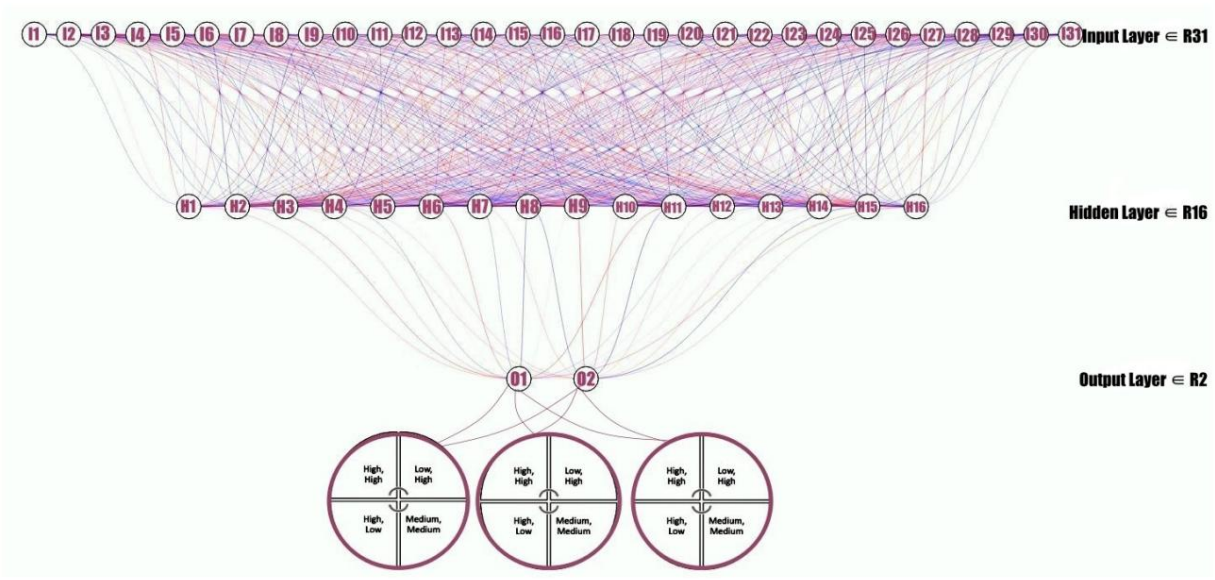


Figura 3. Red neuronal de tres capas para tareas de contabilidad y presentación de informes financieros.
Figura 3. Red neuronal de tres capas para tareas de contabilidad y presentación de informes financieros.

La capa oculta es donde la red neuronal aprende representaciones complejas y La capa oculta es donde
la red neuronal aprende representaciones complejas y
patrones de las entidades de entrada. Al conectar la capa de entrada a la capa oculta, se crean los patrones de
las entidades de entrada. Al conectar la capa de entrada a la capa oculta, la
La red puede extraer conocimientos y relaciones significativas. Si el objetivo de la entidad económica
es aumentar el valor del negocio, tiene sentido correlacionar la capa oculta con la financiera
KPI. Los KPI financieros a menudo sirven como indicadores de desempeño que impactan directamente el resultado general.

Valor y éxito de un negocio. La capa oculta puede aprender a abstraer y combinar características de la capa de entrada para identificar patrones asociados con resultados financieros exitosos o un mayor valor comercial. La elección de las dos capas de producción depende de los objetivos específicos y de cómo la empresa define y se relaciona con el valor del negocio, ya sea por rentabilidad o crecimiento de ingresos. Los KPI exactos pueden variar según la industria, los objetivos de la empresa y la naturaleza del negocio que integra la herramienta de inteligencia artificial de red neuronal.
Uno de los indicadores clave de un negocio exitoso es su capacidad para generar ingresos crecientes con el tiempo. La primera capa de resultados podría representar el crecimiento de ingresos previsto. Luego, la red neuronal se entrenaría para aprender patrones en los datos de entrada que se correlacionen con mayores ingresos.

El modelo matemático tendría dos nodos de salida, cada uno correspondiente a uno de estos resultados. Para la clasificación binaria, la empresa podría utilizar una función de activación sigmoidea en la capa de salida, y para tareas de regresión, la empresa podría utilizar la activación lineal [17].

$$(1)\ Y1 = \sigma(H \cdot W1\ (2) + b1\ (2))$$

$$(2)\ Y2 = \sigma(H \cdot W2\ (2) + b2\ (2))$$

donde σ es la función de activación sigmoidea, e Y1 e Y2 representan los valores previstos para el crecimiento de los ingresos y la rentabilidad, respectivamente.

El tercer nodo de salida se puede incluir si la entidad económica puede tener en cuenta los KPI financieros negativos que desea minimizar, como los costos operativos o el índice de gastos.

$$(3)\ Y3 = \sigma(H \cdot W3\ (2) + b3\ (2))$$

4.2. Uso de herramientas de inteligencia artificial de teoría de juegos en el proceso de toma de decisiones

La teoría de juegos es poderosa en la era de la digitalización [20] y la IA porque proporciona un enfoque estructurado para la toma de decisiones en entornos complejos, dinámicos e inciertos. Ayuda a los tomadores de decisiones a navegar interacciones estratégicas, asignar recursos de manera eficiente y adaptarse al panorama rápidamente cambiante de la tecnología y los negocios. Como mencionamos anteriormente , alentamos a las entidades económicas a aplicar la teoría de juegos a sus procesos comerciales para desbloquear el valor comercial a través de la toma de decisiones impulsada por la IA y la transformación digital en los sistemas de informes financieros y crear una matriz de beneficios que represente las interacciones entre diferentes procesos de toma de decisiones. responsables o partes interesadas. En el contexto de nuestra investigación, estos tomadores de decisiones podrían incluir la dirección de la empresa, los accionistas, el sistema de inteligencia artificial y otras entidades relevantes.

Suponemos que hay dos tomadores de decisiones clave: la dirección de la empresa (CM) y los accionistas (SH), y hay dos opciones estratégicas para cada uno: implementar informes financieros (AI) impulsados por IA o ceñirse a los informes tradicionales (TR). Los beneficios se expresan en términos de valor o utilidad comercial para cada combinación de opciones.

Explicación de los beneficios •

Alto, alto: si tanto la dirección de la empresa como los accionistas optan por implementar informes financieros basados en IA, ambos reciben un alto valor o utilidad comercial. • Bajo, alto: si la

dirección de la empresa opta por la IA mientras los accionistas se atienen a los informes tradicionales, la dirección de la empresa podría experimentar un valor bajo (debido a los costos de implementación, por ejemplo), mientras que los accionistas reciben un valor alto (ya que pueden preferir el sistema tradicional familiar). informes).

• Alto, Bajo: Si la Gerencia de la Compañía se apega a los informes tradicionales mientras que los Accionistas prefieren los informes impulsados por IA, la Gerencia de la Compañía podría lograr un alto valor (ya que evitan los costos de implementación), pero los Accionistas reciben un valor bajo (ya que desean los beneficios de la IA).

informes impulsados). • Medio, Medio: Si tanto la Gerencia de la Compañía como los Accionistas se apegan a los informes tradicionales, ambos logran un nivel medio de valor o utilidad comercial.

Podemos detallar los beneficios en términos de aumento de valor tanto para la dirección de la empresa como para los accionistas basándose en dos opciones estratégicas: implementar informes financieros basados en IA o ceñirse a los informes tradicionales. Los valores representan el valor o utilidad empresarial percibido, y los valores más altos indican un valor percibido más alto. Alto, medio y bajo son representaciones cualitativas del valor percibido, y los valores numéricos reales dependerían del contexto específico y las preferencias de las partes interesadas.

Los beneficios se describen en la Tabla 1, y los valores reales dependerían de factores como la industria, los objetivos y preferencias específicos de las partes interesadas y los beneficios y desventajas percibidos de los informes basados en IA versus los informes tradicionales en el contexto dado. Esta matriz proporciona una forma estructurada de comprender cómo las decisiones de cada parte influyen en el valor percibido tanto por la Dirección de la Empresa como por los Accionistas:

Tabla 1. Matriz de pagos.

		Accionistas	
		AI	TR
Compañía Gestión	AI	Alta alta	Bajo, alto
	TR	Alta baja	Medio, Medio

- Alto, Alto (AI): Gestión de la empresa (CM): Alto valor, ya que se espera que los informes financieros impulsados por IA mejoren la eficiencia, la precisión y la toma de decisiones estratégicas, lo que conducirá a un mayor rendimiento empresarial general.
  Accionistas (SH): Alto valor, ya que se benefician de una mayor transparencia, una toma de decisiones mejor informada por parte de la gerencia y ganancias potencialmente mayores.
- Bajo, Alto (TR): Gestión de la empresa (CM): valor bajo, ya que apegarse a los informes tradicionales puede resultar en la pérdida de oportunidades de ganancias de eficiencia, conocimientos estratégicos y ahorros de costos que ofrecen los informes impulsados por IA.
  Accionistas (SH): Alto valor, ya que pueden preferir la familiaridad y estabilidad de los informes tradicionales, percibiendo potencialmente menos riesgo o interrupción de sus inversiones.
- Alto, Bajo (AI): Gestión de la empresa (CM): Alto valor, ya que la implementación de informes impulsados por IA satisface el objetivo de la dirección de adoptar tecnologías innovadoras y mantenerse competitivo.
  Accionistas (SH): Valor bajo, ya que podrían sentirse decepcionados con la decisión de no seguir con los informes tradicionales, percibiendo potencialmente mayores riesgos o incertidumbres.
- Medio, Medio (TR):
  Gestión de la empresa (CM): valor medio, ya que apegarse a los informes tradicionales puede proporcionar estabilidad pero no aprovechar los beneficios potenciales que ofrecen los informes basados en IA.
  Accionistas (SH): Valor medio, ya que mantienen una sensación de estabilidad pero pueden perderse posibles mejoras en la toma de decisiones y la eficiencia.

Los directores financieros enfrentan tanto desafíos como oportunidades en el proceso de toma de decisiones de IA dentro de un entorno industrial y regulatorio determinado. Estos pueden variar según las circunstancias específicas de cada organización. Deben navegar por complejas regulaciones de privacidad de datos para garantizar que el uso de la IA se alinee con las leyes de privacidad, particularmente en industrias que manejan información confidencial de los clientes. Otro desafío que la gerencia enfrentará pronto es mantenerse al día con las regulaciones en evolución relacionadas con la IA y garantizar que los sistemas de IA cumplan con las leyes y estándares específicos de la industria.

Las matrices de la teoría de juegos se pueden adaptar a diversos escenarios de toma de decisiones y, cuando se aplican al contexto de los directores financieros (CFO) que toman decisiones en el ámbito de la IA, proponemos cuatro matrices que se pueden tener en cuenta en el proceso de toma de decisiones. : Los directores financieros que toman decisiones relacionadas con la IA pueden utilizar matrices de teoría de juegos para navegar en cuatro escenarios clave : adopción temprana versus tardía de la IA, desarrollo de IA de proveedores externos versus desarrollo interno, colaboración para compartir datos versus protección de datos y cumplimiento regulatorio proactivo versus reactivo. , equilibrando beneficios, costos y riesgos para lograr ventajas competitivas e innovación.

4.3. Cuestiones éticas y de privacidad en los informes financieros basados en IA

En este entorno relativamente nuevo de informes financieros impulsado por la IA, las preocupaciones sobre la privacidad de los datos son primordiales. Los amplios requisitos de datos de los sistemas de IA a menudo abarcan información financiera y personal confidencial, lo que plantea el espectro de acceso no autorizado y violaciones de datos [27]. La base misma de nuestro trabajo se basa en la integridad y confidencialidad de los datos que manejamos. A medida que integramos la IA, debemos estar atentos a la protección de estos datos mediante métodos de cifrado sólidos, garantizando que la información confidencial permanezca segura tanto en tránsito como en reposo. Las partes interesadas deben estar plenamente informadas sobre cómo se recopilan, almacenan y utilizan sus datos. Esta transparencia se extiende a la obtención del consentimiento explícito de las personas cuyos datos se utilizan. Al implementar políticas estrictas de gobernanza de datos, podemos delinear controles de acceso claros y garantizar que solo el personal autorizado pueda acceder a los datos en circunstancias apropiadas. La rendición de cuentas y la transparencia en los sistemas de IA son fundamentales y es imperativo que los desarrolladores de aplicaciones establezcan marcos de rendición de cuentas claros que definan las responsabilidades de los usuarios, operadores y tomadores de decisiones de IA.
Cuando ocurren errores, debe haber un proceso transparente para identificar y abordar el origen del problema. La implementación de técnicas de IA explicables puede ayudar significativamente en este proceso. Estas técnicas nos permiten comprender y articular cómo los sistemas de IA toman decisiones, fomentando así la confianza y la responsabilidad entre las partes interesadas [28].

Con la adopción generalizada de normas contables internacionales, el cumplimiento normativo es una preocupación siempre presente en nuestra profesión. Las aplicaciones de IA en los informes financieros deben cumplir con una gran variedad de requisitos regulatorios, desde regulaciones financieras hasta leyes de protección de datos. Participar en un diálogo continuo con los organismos reguladores puede proporcionar una orientación valiosa y ayudar a garantizar que nuestros sistemas de IA cumplan con las leyes vigentes. Las auditorías de cumplimiento periódicas son esenciales para verificar el cumplimiento y abordar cualquier brecha en nuestros procesos.

Los riesgos de seguridad son un aspecto inherente a la integración de la IA en los informes financieros. Los sistemas de IA pueden ser blanco de ciberataques, lo que podría provocar filtraciones de datos o manipulación de información financiera [29]. Para mitigar estos riesgos, debemos implementar medidas avanzadas de ciberseguridad y desarrollar planes integrales de respuesta a incidentes. Las evaluaciones de seguridad periódicas ayudarán a identificar y abordar las vulnerabilidades, garantizando la integridad y confiabilidad de nuestros sistemas de inteligencia artificial. En situaciones de la vida real, varias medidas avanzadas de ciberseguridad han demostrado ser efectivas para salvaguardar los sistemas de informes financieros impulsados por IA, como la autenticación multifactor que mejora significativamente la seguridad al exigir a los usuarios que proporcionen dos o más factores de verificación para obtener acceso a un sistema. , cifrado de extremo a extremo que garantiza que los datos estén cifrados desde el momento en que se crean hasta que el destinatario previsto los recibe y los descifra, o realizar auditorías de seguridad periódicas y pruebas de penetración que permitan a las organizaciones identificar y abordar vulnerabilidades en sus sistemas de forma proactiva [29].

5. Resultados y Discusión

Combinar un modelo de red neuronal con la teoría de juegos, representada en una matriz, puede ser un enfoque poderoso, y a menudo se lo conoce como aprendizaje automático de teoría de juegos y se utiliza para respaldar las decisiones en las empresas. La red neuronal se puede utilizar para predecir KPI financieros u otros resultados relevantes, y la matriz de teoría de juegos puede ayudar a analizar las interacciones estratégicas entre diferentes entidades o partes interesadas. Pretende comprender cómo diferentes agentes o jugadores, cada uno con sus propios objetivos, toman decisiones en un entorno competitivo o

ambiente cooperativo. Utilizamos la teoría de juegos como una rama de las matemáticas y la economía para monitorear las interacciones entre tomadores de decisiones racionales, a menudo mencionadas en la literatura como jugadores, en situaciones donde el resultado de la decisión de un jugador depende del decisiones de otros, y proponemos combinar técnicas de aprendizaje automático e integrar ellos con la teoría de juegos para modelar y predecir el comportamiento de los jugadores que pueden beneficiarse de creación positiva de valor empresarial. En nuestra teoría de juegos, los jugadores representan a los Accionistas. y el equipo directivo, y las estrategias incluyen una mayor capitalización de mercado mediante aumentar los ingresos o la rentabilidad y los pagos representados en la matriz. Estrategias propuestas por la investigación son las opciones disponibles para cada jugador, como elegir el Digitalización e incorporación de IA o apego a métodos tradicionales de contabilidad. y los informes, y los pagos representan los resultados asociados con combinaciones específicas de estrategias elegidas por todos los jugadores. Para aplicar las estrategias se utilizan modelos predictivos, como el Las redes neuronales con capas de árbol se pueden emplear para estimar las elecciones o acciones probables. de jugadores en base a datos contables históricos extraídos del balance y del beneficio y pérdida de datos u otras características relevantes.

Dadas las especificaciones de nuestra red neuronal y la estructura de la matriz de pagos Desde la teoría de juegos, sugerimos nombrar el modelo teórico “Decision Harbor AI”. que refleja la integración de la inteligencia artificial, la toma de decisiones financieras y las estrategias análisis y simboliza un puerto seguro e informado para la toma de decisiones con la ayuda de inteligencia. El aprendizaje automático basado en la teoría de juegos proporciona un marco para la comprensión y predecir interacciones estratégicas [30]. Al integrar modelos de aprendizaje automático con conceptos de la teoría de juegos, predecimos que será posible obtener conocimientos sobre los procesos de toma de decisiones en entornos complejos y dinámicos. Este enfoque puede ser valioso para tomar decisiones informadas en escenarios competitivos o cooperativos donde múltiples Las partes interesadas están involucradas [31].

En nuestro escenario, tenemos una matriz de teoría de juegos con dos jugadores: la dirección de la empresa y los accionistas, y la matriz representa los posibles resultados en función de la decisiones tomadas por estos jugadores. Podemos entrenar la red neuronal para predecir las finanzas. KPI o resultados de interés, como crecimiento de ingresos, rentabilidad y eficiencia de costos, como como se ve en la Figura 4. Después de entrenar la red neuronal para predecir indicadores de desempeño positivos , usamos la red neuronal entrenada para hacer predicciones para cada jugador (Empresa Management, AI y TR) basados en las características de entrada y el KPI en la capa oculta.

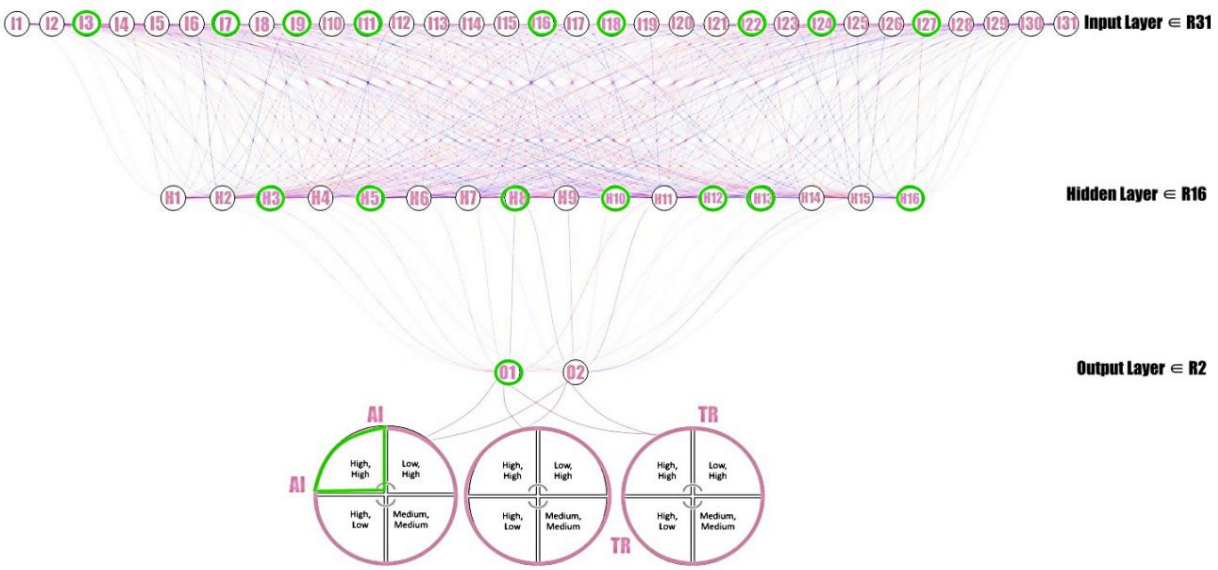


Figura 4. Modelo teórico de IA de Decision Harbor: escenario 1.

Figura 4. Modelo teórico de IA de Decision Harbor: escenario 1.

En los malos años financieros, la distribución de los datos puede cambiar. Los patrones y las relaciones entre las características de los insumos y los resultados financieros pueden cambiar, lo que lleva a una falta de coincidencia entre los datos de capacitación y de prueba. La red neuronal, entrenada con datos históricos, puede tener dificultades para generalizarse a estas nuevas condiciones. El rendimiento predictivo del modelo puede degradarse durante malos años financieros si los patrones observados durante el entrenamiento ya no se mantienen. El modelo podría realizar predicciones inexactas, especialmente si no ha encontrado escenarios similares en los datos de entrenamiento (Figura 5).

Figura 4. Modelo teórico de IA de Decision Harbor: escenario 1.

En los malos años financieros, la distribución de los datos puede cambiar. Los patrones y las relaciones
En los malos años financieros, la distribución de los datos puede cambiar. Los patrones y la relación entre
las características de los insumos y los resultados financieros pueden cambiar, lo que lleva a un desajuste entre
La relación entre las características de entrada y los resultados financieros puede cambiar, lo que lleva a una
discrepancia entre los datos de capacitación y prueba. La red neuronal, entrenada con datos históricos, puede tener dificultades
entre los datos de entrenamiento y de prueba. La red neuronal, entrenada con datos históricos, puede generalizarse a
estas nuevas condiciones. El rendimiento predictivo del modelo puede degradarse
Luchamos por generalizar a estas nuevas condiciones. El rendimiento predictivo del modelo puede verse afectado
durante malos años financieros si los patrones observados durante el entrenamiento ya no se mantienen. El
degradarse durante los malos años financieros si los patrones observados durante el entrenamiento ya no pueden
hacer predicciones inexactas, especialmente si no se han encontrado similares
sostener. El modelo podría realizar predicciones inexactas, especialmente si no ha encontrado
escenarios en los datos de entrenamiento (Figura 5).
escenarios similares en los datos de entrenamiento (Figura 5).

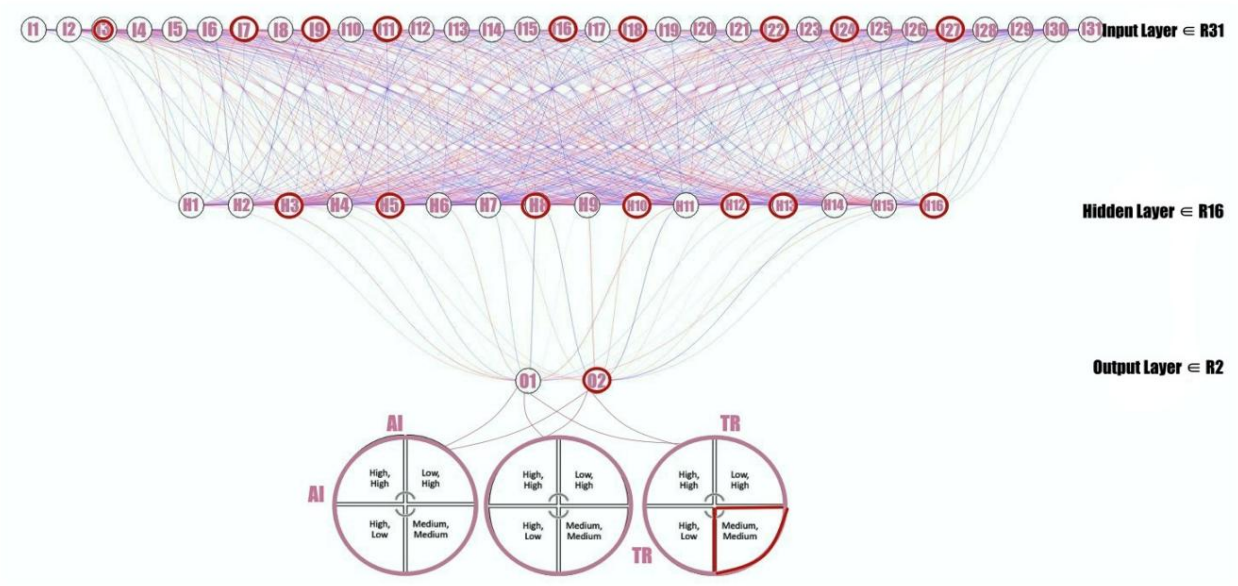


Figura 5. Modelo teórico de IA de Decision Harbor: escenario 2.
Figura 5. Modelo teórico de IA de Decision Harbor: escenario 2.

Sugerimos que la empresa implemente una estrategia de capacitación continua de modelos Sugerimos que la
empresa implemente una estrategia de capacitación continua de modelos para
adaptarse a la evolución de las condiciones financieras. El modelo debe actualizarse periódicamente para adaptarse a
la evolución de las condiciones financieras. El modelo debe actualizarse periódicamente con nuevas
nuevos datos, especialmente durante períodos de turbulencia financiera, para garantizar que sigan siendo relevantes.
datos, especialmente durante períodos de turbulencia financiera, para garantizar que sigan siendo relevantes. Para
Para que el modelo conduzca a buenas decisiones financieras, la dirección debe llevar a cabo un escenario.
que el modelo conduzca a buenas decisiones financieras, la dirección debe llevar a cabo un escenario.
Análisis pario para evaluar el desempeño del modelo bajo diversas condiciones financieras. Este análisis sirve para
evaluar el rendimiento del modelo en diversas condiciones financieras. Esto puede
puede implicar probar el modelo contra escenarios simulados que representan diferentes economías .
estados económicos, incluidos los buenos y malos ejercicios financieros. Si la red neuronal funciona bien , se incluyen
estados financieros buenos y malos. Si la red neuronal funciona bien durante
buenos ejercicios financieros, proporciona una validación de la solidez del modelo. Sugiere que
El modelo puede adaptarse a diferentes condiciones financieras y continúa siendo preciso.
predicciones cuando las condiciones son favorables. Los buenos ejercicios financieros pueden reforzar positivamente
el uso del modelo para la toma de decisiones. Predicciones confiables en nuestra empresa durante
Estos períodos pueden contribuir a una planificación estratégica efectiva, asignación de recursos y otros
decisiones que mejoran el valor del negocio. Los tomadores de decisiones pueden ganar mayor confianza
en las predicciones del modelo durante buenos ejercicios financieros. Esta confianza puede conducir a más
Confianza en el modelo para guiar las decisiones.

El ciclo de Deming (PDCA) se puede utilizar como un método de gestión iterativo de cuatro pasos.
para la mejora continua. Aplicar el ciclo PDCA a la formación continua y
Análisis de escenarios de nuestro modelo de red neuronal en el contexto de condiciones financieras en evolución.
es un enfoque sensato y eficaz. Debemos monitorear el proceso de formación y evaluar
Si el rendimiento del modelo mejora o permanece estable, ejecute el análisis de escenario.
utilizando los scripts o herramientas de simulación preparados y evaluar la precisión, sensibilidad,
y especificidad en diferentes escenarios para identificar cualquier patrón de bajo rendimiento
o sobreajuste.

Si el análisis de escenarios de nuestra red neuronal revela deficiencias, debemos considerar
ajustar el modelo o el proceso de generación de escenarios. Esto puede implicar redefinir características ,
introducir nuevas características o modificar el enfoque de simulación. El teórico

El modelo puede incorporar cualquier mejora necesaria para mejorar la capacidad del modelo para manejar diversas condiciones financieras. Podemos aplicar pruebas de tensión a nuestro modelo [23], como la forma en que los bancos realizan pruebas de tensión para evaluar la resiliencia de sus sistemas financieros en condiciones adversas. En el contexto de nuestro modelo de red neuronal y nuestra matriz de teoría de juegos, las pruebas de estrés pueden ayudarnos a comprender qué tan bien se desempeña el modelo y qué tan sólida es la toma de decisiones frente a escenarios financieros extremos o inesperados. El aprendizaje profundo se puede aplicar a la capa de entrada de nuestra red en el contexto de las pruebas de estrés dinámicas del balance [31,32]. Si el balance que genera información financiera para los datos de la capa de entrada tiene una dimensión temporal como trimestral o anual, podemos considerar redes neuronales recurrentes o redes de memoria a corto plazo (LSTM) para capturar patrones temporales y dependencias en los datos. Si nuestros datos de entrada incluyen diversas fuentes como estados financieros, datos de mercado e indicadores económicos, sugerimos aplicar arquitecturas que admitan la integración de datos multimodales. Esto permite que el modelo aprenda de diferentes tipos de información simultáneamente. En general, es importante adaptar el enfoque de aprendizaje profundo a los requisitos y características específicos de nuestros datos financieros y objetivos de pruebas de estrés. Nuestro modelo teórico DecisionHarborAI también se puede implementar en organizaciones públicas . La aplicación de herramientas de inteligencia artificial, como este modelo teórico de dos capas, en el sector público tiene el potencial de generar beneficios significativos, como ayudar a analizar el impacto de diferentes políticas, ayudar a los formuladores de políticas a comprender los resultados potenciales y tomar decisiones informadas, proporcionando información para optimizar asignación de recursos para lograr los resultados deseados, evaluar la eficacia de los programas e iniciativas públicas y mejorar la participación ciudadana proporcionando información basada en datos sobre los procesos de toma de decisiones, fomentando la transparencia y la confianza entre el público y el gobierno [33].

Para implementar efectivamente tal modelo, surgen varias recomendaciones prácticas. Las actualizaciones periódicas y el reentrenamiento de la red neuronal con nuevos datos son cruciales para mantener la precisión y la relevancia.

El primer paso es proporcionar una imagen o un archivo PDF de un balance a nuestra API. Esto se llevará a cabo en una aplicación móvil o una aplicación web. Tan pronto como se recibe una imagen o un PDF, se convierte en un archivo TXT. Un analizador toma el TXT obtenido del OCR y lo convierte en JSON estructurado mediante el aprendizaje automático. Luego, el JSON se devuelve como salida de la API. Desde aquí se pueden seguir procesando los datos del balance. Luego, el conjunto de datos completo se utiliza para entrenar una red neuronal, capaz de predecir indicadores clave de desempeño financiero, como el crecimiento de los ingresos, la rentabilidad y la eficiencia de costos. El modelo DecisionHarborAI va un paso más allá al incorporar un marco de teoría de juegos. Esto permite un análisis matizado de las interacciones estratégicas entre las partes interesadas, como la administración y los accionistas. Al comprender estas dinámicas, la institución puede alinear sus decisiones estratégicas con los intereses de todas las partes involucradas, maximizando así el valor comercial.

Para garantizar la solidez, se realizan análisis continuos de escenarios y pruebas de estrés. Esta práctica no sólo valida las predicciones del modelo en diferentes condiciones sino que también prepara a la institución para posibles fluctuaciones económicas. Los resultados son profundos: capacidades mejoradas para la toma de decisiones, ganancias significativas de eficiencia a través de la automatización y una alineación estratégica que impulsa el desempeño empresarial general.

También se deben considerar las limitaciones al implementar modelos de informes financieros basados en IA, como DecisionHarborAI. Una de las limitaciones más importantes es la calidad de los datos introducidos en los modelos de IA. Los sistemas de IA requieren grandes cantidades de datos de alta calidad para funcionar de manera óptima. Sin embargo, los datos financieros a menudo pueden ser incompletos, inconsistentes o confusos. Los datos inexactos o de mala calidad pueden dar lugar a predicciones erróneas y procesos de toma de decisiones defectuosos. Las organizaciones deben invertir un esfuerzo considerable en la limpieza y validación de datos para garantizar que los datos utilizados con fines operativos y de capacitación sean precisos y confiables. Esto puede requerir muchos recursos y no siempre es factible, particularmente para instituciones más pequeñas con presupuestos limitados. Como todos los modelos de IA, DecisionHarborAI se entrenará con datos históricos y puede funcionar bien en condiciones similares a las presentes en el conjunto de datos de entrenamiento. La capacidad de nuestro modelo teórico para adaptarse a los cambios económi

Los cambios regulatorios y los eventos globales imprevistos, como crisis financieras o pandemias, son una preocupación crítica [34]. Es posible que el modelo DecisionHarborAI no se generalice de manera efectiva a escenarios nuevos e invisibles, lo que lleva a una degradación del rendimiento. Será necesario un reentrenamiento y validación continuos del modelo para garantizar que DecisionHarborAI siga siendo relevante y preciso, pero esto puede ser un desafío de administrar y requiere una inversión continua.

6. Conclusiones

Al implementar una red neuronal para la contabilidad financiera, es importante considerar factores como la calidad de los datos, la interpretabilidad del modelo y el entorno regulatorio específico. El modelo que propone esta investigación ofrece un enfoque integrado para el apoyo a la toma de decisiones al combinar las capacidades predictivas de las redes neuronales con los conocimientos estratégicos proporcionados por la teoría de juegos. Esta integración tiene como objetivo mejorar los procesos de toma de decisiones en entornos complejos y dinámicos y cambia el enfoque del modelo para predecir indicadores financieros y respaldar la toma de decisiones financieras. El componente de red neuronal de la nueva herramienta de IA está diseñado para proporcionar predicciones e información financiera basadas en datos históricos. Esto incluye pronosticar indicadores financieros clave, como el crecimiento de los ingresos, la rentabilidad y la eficiencia de costos. Las redes neuronales pueden ser una herramienta valiosa para calibrar resultados en el entorno de la contabilidad financiera y han demostrado eficacia en el procesamiento de datos financieros complejos, la identificación de patrones y la realización de predicciones . La inclusión de una matriz de resultados de la teoría de juegos en esta nueva herramienta de toma de decisiones de IA introduce un marco estratégico para analizar las interacciones entre los tomadores de decisiones.
Esto permite considerar opciones y resultados estratégicos en un contexto más dinámico y competitivo. Las herramientas de transformación digital de IA tienen el potencial de mejorar significativamente los procesos de toma de decisiones financieras y la gestión estratégica general. Estas herramientas pueden ayudar a las organizaciones a aprovechar conocimientos basados en datos y automatizar tareas rutinarias, mejorando potencialmente la eficiencia y la precisión de las decisiones. Sin embargo, es importante reconocer que las tecnologías de IA también presentan desafíos y desventajas, como preocupaciones sobre la privacidad de los datos, el riesgo de sesgos algorítmicos y la necesidad de importantes recursos computacionales. Además, la complejidad de los modelos de IA puede hacer que sean difíciles de interpretar y de confiar, lo que puede limitar su aplicación práctica en determinados contextos. Por lo tanto, si bien la IA ofrece avances prometedores, es esencial abordar su integración con una perspectiva equilibrada, reconociendo tanto sus beneficios como sus limitaciones. Todas las herramientas de asistencia a la toma de decisiones que se desarrollen en el futuro deben aprender y actualizar continuamente sus predicciones, garantizando relevancia y precisión a lo largo del tiempo. El modelo que sugiere este artículo debe estar equipado con la capacidad de someterse a pruebas de estrés y análisis de escenarios, que le permitan evaluar su desempeño bajo diversas condiciones financieras, incluidos escenarios tanto favorables como adversos, lo que contribuya a su solidez.

La fusión de matrices de pagos de redes neuronales representa una herramienta integral y adaptable de apoyo a las decisiones que aprovecha las fortalezas de las redes neuronales y la teoría de juegos para proporcionar información valiosa para la toma de decisiones financieras en los sectores público y privado y proporciona un valor tangible a las empresas. La capacitación continua, el análisis de escenarios y el marco estratégico contribuyen a su potencial efectividad en entornos dinámicos e inciertos. El concepto de integrar la toma de decisiones impulsada por la IA se refleja directamente en las conclusiones de este documento. Las herramientas digitales combinadas combinan las capacidades predictivas de las redes neuronales con un marco de teoría de juegos, enfatizando la integración de técnicas avanzadas de inteligencia artificial para una toma de decisiones más informada. El éxito del modelo dependerá de pruebas exhaustivas, la colaboración con expertos en el campo y mejoras continuas basadas en comentarios del mundo real.

Las herramientas de transformación digital pueden ayudar significativamente en el proceso de toma de decisiones financieras y en la gestión estratégica general de una empresa al analizar datos financieros históricos para hacer predicciones precisas sobre tendencias y resultados futuros. Estas herramientas pueden evaluar y predecir riesgos potenciales mediante el análisis de diversas fuentes de datos. Esto permite a las organizaciones identificar y mitigar proactivamente los riesgos en la toma de decisiones financieras y automatizar tareas repetitivas y rutinarias, como la entrada de datos, la conciliación y la generación de informes.

Las herramientas de transformación digital de IA desempeñan un papel crucial en la mejora de los procesos de toma de decisiones financieras y la gestión estratégica general. Estas herramientas permiten a las organizaciones aprovechar conocimientos basados en datos, automatizar tareas rutinarias y navegar por las complejidades del panorama empresarial moderno de forma más eficaz.

Contribuciones de los autores: Conceptualización, AEA y AED; AED y LI: análisis formal, AEA: investigación, LI: recursos, AED: redacción—preparación del borrador original, AEA: redacción—revisión y edición, AED: visualización, LI: supervisión, AEA: administración del proyecto. Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: Esta investigación no recibió financiamiento externo.

Declaración de disponibilidad de datos: todos los datos subyacentes a los resultados están disponibles como parte del artículo y no se requieren datos de origen adicionales.

Conflictos de intereses: Los autores declaran no tener conflictos de intereses.

Apéndice A

Tabla A1. Descripción de las capas en el modelo neuronal propuesto: Decision Harbor AI.

Capa de entrada R^{31} I	Capa oculta R ····· h	Capa de salida R^{2} oh
Ingresos netos	Ingresos brutos	Crecimiento de ingresos
Lngresos netos		
Impuestos	Ingresos brutos	
Gastos por intereses	EBITDA	
Depreciación		
Amortización	Márgenes de beneficio	
Costo de los bienes vendidos		
Los gastos de explotación	Flujo de caja de operaciones	
Ingresos de explotación		
Cambio en el capital de trabajo	Flujo de caja libre	
Flujo de caja operativo		
Gastos de capital	HUEVA	
Accionistas		
Activos totales promedio	ROA	
Responsabilidad total		
Activos circulantes	Relación deuda-capital	
Pasivo circulante		
Dinero	Radio actual	Rentabilidad
Equivalentes de efectivo		
cuentas por cobrar netas	Razón rápida	
Gastos pagados por anticipado		
Precio de la acción	Relación precio/beneficios	
Ganancias por acción		
Capital preferente	Libro evaluado por compartir	
Dividendo por acción		
Valor por acción	Rentabilidad por dividendo	
Acciones totales		
Crecimiento de ingresos	Capitalización de mercado	
Ganancias		
voluntad divina	Métricas de crecimiento	
Activos fijos		

Referencias

1. Imoniana, JO; Reginato, L.; Júnior, EBC; Benetti, C. Percepciones de las partes interesadas sobre la adopción de Normas Internacionales de Información Financiera por parte de las pequeñas y medianas empresas: un análisis multidimensional. En t. J. Autobús. Emergente. Marca. 2025, 1. [Referencia cruzada]
2. Radicic, D.; Petkovi´c, S. Impacto de la digitalización en las innovaciones tecnológicas en las pequeñas y medianas empresas (PYMES). Tecnología. Pronóstico. Soc. Chang. 2023, 191, 122474. [Referencia cruzada]
3. Saura, JR; Ribeiro-Soriano, D.; Palacios-Marqués, D. Evaluación de las cuestiones de privacidad de la ciencia de datos conductuales en el gobierno artificial despliegue de inteligencia. Gobernador Inf. P. 2022, 39, 101679. [CrossRef]
4. Faccia, A.; Al Naqbi, MYK; Lootah, SA Ciclo Integrado de Contabilidad Financiera en la Nube. En Actas de la Tercera Conferencia Internacional sobre Computación en la Nube y Big Data de 2019, Oxford, Reino Unido, 28 a 30 de agosto de 2019; ACM: Nueva York, NY, EE. UU., 2019; págs. 31–37. [Referencia cruzada]
5. Allam, K.; Rodwal, A. Análisis de big data impulsado por IA: revelación de conocimientos para el avance empresarial. EPH-Int. J. Ciencias. Ing. 2023, 9, 53–58. [Referencia cruzada]
6. Mengu, D.; Luo, Y.; Rivenson, Y.; Ozcan, A. Análisis de redes neuronales ópticas difractivas y su integración con electrónica Redes neuronales. IEEE J. Sel. Arriba. Electrón cuántico. 2020, 26, 1-14. [Referencia cruzada]
7. Shengelia, NSN; Tsiklauri, ZTZ; Rzepka, ARA; Shengelia, RSR El impacto de las tecnologías financieras en la transformación digital mación de Contabilidad, Auditoría e Información Financiera. Economía 2022, 105, 385–399. [Referencia cruzada]
8. Pacurari, D.; Nechita, E. Algunas consideraciones sobre la contabilidad en la nube. En Estudios e Investigaciones Científicas. Edición de Economía; Facultad de Ciencias Económicas, Universidad “Vasile Alecsandri” de Bacau: Bacau, Rumania, 2013. [CrossRef]
9. Li, M.; Yu, Y.; Li, X.; Zhao, JL; Zhao, D. Determinantes de la transformación de las PYMES hacia servicios en la nube: perspectivas de racionalidades económicas y sociales. Pac. Asia J. Asociación. Inf. Sistema. 2019, 11, 65–87. [Referencia cruzada]
10. Sorensen, M.; Yasuda, A. Impacto de las inversiones de capital privado en las partes interesadas. En Manual de economía de las finanzas corporativas: capital privado y finanzas empresariales; Elsevier: Ámsterdam, Países Bajos, 2023; págs. 299–341. [Referencia cruzada]
11. Naik, N.; Hameed, BZ; Shetty, DK; Swain, D.; Shah, M.; Pablo, R.; Aggarwal, K.; Ibrahim, S.; Patil, V.; Smriti, K.; et al. Consideración jurídica y ética de la inteligencia artificial en la atención sanitaria: ¿quién asume la responsabilidad? Frente. Cirugía. 2022, 9, 862322. [Referencia cruzada]
12. Vo-Thanh, T.; Zamán, M.; Hasán, R.; Akter, S.; Dang-Van, T. La digitalización del servicio en restaurantes de alta cocina: un costo-beneficio perspectiva. En t. J. Contemporáneo. Hosp. Gestionar. 2022, 34, 3502–3524. [Referencia cruzada]
13. Agustí-Juan, I.; Vidrio, J.; Pawar, V. Un cuadro de mando integral para evaluar la automatización en la construcción. En Actas de la Conferencia sobre Construcción Creativa 2019, Budapest, Hungría, 29 de junio a 2 de julio de 2019; Universidad de Tecnología y Economía de Budapest: Budapest, Hungría, 2019; págs. 155-163. [Referencia cruzada]
14. Wang, Z.; Wang, Q.; Nie, Z.; Li, B. Predicción de dificultades financieras corporativas basada en la promesa de capital de los accionistas mayoritarios. Aplica. Economía. Letón. 2022, 29, 1365–1368. [Referencia cruzada]
15. Du, G.; Elston, F. ARTÍCULO RETRACTADO: Evaluación de riesgos financieros para mejorar la precisión de la predicción financiera en la industria financiera de Internet utilizando modelos de análisis de datos. Ópera. Gestionar. Res. 2022, 15, 925–940. [Referencia cruzada]
16. Canción, Z.; Wu, S. Post teoría de juegos de previsión financiera y toma de decisiones. Finanzas. Res. Letón. 2023, 58, 104288. [Referencia cruzada]
17. Murali, M.; Kanmani, S. Modelado de aplicaciones de teoría de juegos utilizando elementos de red para garantizar el equilibrio de Nash a través de Price de la anarquía y el precio de la estabilidad. En t. J. Sistema. Sistema. Ing. 2024, 14. [Referencia cruzada]
18. Hamido¨glu, A. Un enfoque de teoría de juegos sobre la construcción de un nuevo modelo de cadena de suministro agroalimentario apoyado por el gobierno. Experto. Sistema. Aplica. 2024, 237, 121353. [Referencia cruzada]
19. Zhuang, YT; Wu, F.; Chen, C.; Pan, YH Retos y oportunidades: Del big data al conocimiento en IA 2.0. Frente. Inf. Tecnología. Electrón. Ing. 2017, 18, 3-14. [Referencia cruzada]
20. Mahmood, A.; Al Marzooqi, A.; El Khatib, M.; AlAmeemi, H. Cómo la inteligencia artificial puede aprovechar el sistema de información para la gestión de proyectos (PMIS) y la toma de decisiones basada en datos en la gestión de proyectos. En t. J. Autobús. Anal. Seguro. (IJBAS) 2023, 3, 184-195. [Referencia cruzada]
21. Chongcs, J.; Katiarayan, V.; Chong, J.; Sin, C. El papel de la inteligencia artificial en las oportunidades, desafíos e implicaciones de la toma de decisiones estratégicas para los gerentes en la era digital. En t. J. Gestionar. Comer. Innovación. 2023, 11, 73–79.
22. Laczkovich, M. Un teorema de superposición del tipo Kolmogorov para funciones continuas acotadas. J. Aprox. Teoría 2021, 269, 105609. [Referencia cruzada]
23. Quraisy, A. Normalidad de los datos mediante las pruebas de Kolmogorov-Smirnov y Shapiro-Wilk. J-HEST J. Educación para la salud. Economía. Ciencia. Tecnología. 2020, 3, 7–11. [Referencia cruzada]
24. Schmidt-Hieber, J. Revisión del teorema de representación de Kolmogorov-Arnold. Red neuronal. 2021, 137, 119-126. [Referencia cruzada]
25. Ismailov, VE Una red neuronal de tres capas puede representar cualquier función multivariada. J. Matemáticas. Anal. Aplica. 2023, 523, 127096. [Referencia cruzada]
26. Elahi, M.; Afolaranmi, SO; Lastra, JLM; García, JAP Una revisión exhaustiva de la literatura sobre las aplicaciones de las técnicas de IA. a lo largo del ciclo de vida de los equipos industriales. Descubrimiento. Artif. Intel. 2023, 3, 43. [Referencia cruzada]
27. Habbal, A.; Ali, MK; Abuzaraida, MA Gestión de confianza, riesgos y seguridad en inteligencia artificial (AI TRiSM): marcos, aplicaciones, desafíos y futuras direcciones de investigación. Experto. Sistema. Aplica. 2024, 240, 122442. [Referencia cruzada]
28. Cheng, L.; Liu, X. De los principios a las prácticas: la interacción intertextual entre los discursos éticos y legales de la IA. En t. J. Pierna. Discurso 2023, 8, 31–52. [Referencia cruzada]

29. Humphreys, D.; Koay, A.; Desmond, D.; Mealy, E. La exageración de la IA como riesgo de seguridad cibernética: la responsabilidad moral de implementar IA generativa en los negocios. Ética de la IA 2024. [CrossRef]
30. Cho, S.; Vasarhelyi, MA; Sol, T.; Zhang, C. Aprendiendo del aprendizaje automático en contabilidad y aseguramiento. J. Emerg. Tecnología. Cuenta. 2020, 17, 1–10. [Referencia cruzada]
31. Ganaie, MA; Hu, M.; Malik, Alaska; Tanveer, M.; Suganthan, PN Aprendizaje profundo de Ensemble: una revisión. Ing. Aplica. Artif. Intel. 2022, 115, 105151. [Referencia cruzada]
32. Zhang, Z.; Cui, P.; Zhu, W. Aprendizaje profundo en gráficos: una encuesta. Traducción IEEE. Conocimiento. Ing. de datos. 2022, 34, 249–270. [Referencia cruzada]
33. Chandana Charitha, P.; Hemaraju, B. Impacto de la inteligencia artificial en la toma de decisiones en las organizaciones. En t. J. Multidisciplinar. Res. 2023, 5. [Referencia cruzada]
34. Aldoseri, A.; Al-Khalifa, KN; Hamouda, AM Repensar la estrategia de datos y la integración de la inteligencia artificial: conceptos, Oportunidades y desafíos. Aplica. Ciencia. 2023, 13, 7082. [Referencia cruzada]

electrónica

MDPI

Artículo

Aprender a mejorar la eficiencia operativa a partir del error de postura Estimación en polinización robótica

Jinlong Chen [1], Jun Xiao 1,*, Minghao Yang [2] y colgar sartén [3]

[1] Escuela de Ciencias de la Computación y Seguridad de la Información, Universidad de Tecnología Electrónica de Guilin, Guilin 541000, China; 7259@guet.edu.cn
[2] Centro de Investigación para la Inteligencia Inspirada en el Cerebro (BII), Instituto de Automatización, Academia China de Ciencias (CASIA), Beijing 100190, China; mhyang@nlpr.ia.ac.cn
[3] Departamento de Ciencias de la Computación, Universidad de Changzhi, Changzhi 046011, China; panhang@czc.edu.cn
* Correspondencia: 22032303110@mails.guet.edu.cn

Resumen: Los robots de polinización autónomos han sido ampliamente discutidos en los últimos años. Sin embargo, la estimación precisa de las poses de las flores en entornos agrícolas complejos sigue siendo un desafío. Con este fin, este trabajo propone la implementación de una arquitectura basada en transformadores para aprender los errores de traslación y rotación entre el efector final del robot de polinización y el objeto objetivo con el objetivo de mejorar la eficiencia de la polinización robótica en tareas de cruzamiento. Las contribuciones son las siguientes: (1) Hemos desarrollado un modelo de arquitectura transformadora, equipada con dos redes neuronales de retroalimentación que hacen retroceder directamente los errores de traslación y rotación entre el efector final del robot y el objetivo de polinización. (2) Además, hemos diseñado una función de pérdida de regresión que se guía por los errores de traslación y rotación entre el efector final del robot y los objetivos de polinización. Esto permite que el brazo robótico identifique de forma rápida y precisa el objetivo de polinización desde la posición actual. (3) Además, hemos diseñado una estrategia para adquirir fácilmente una cantidad sustancial de muestras de entrenamiento a partir de la observación ojo en mano, que pueden utilizarse como entradas para el modelo. Mientras tanto, los errores de traslación y rotación identificados en el sistema de coordenadas cartesianas del manipulador final se designan como objetivos de pérdida simultáneamente. Esto ayuda a optimizar el entrenamiento del modelo. Realizamos experimentos con un sistema de polinización robótico realista. Los resultados demuestran que el método propuesto supera al método más moderno, tanto en términos de precisión como de eficiencia.

Palabras clave: robot polinizador; transformador; errores de compensación

Cita: Chen, J.; Xiao, J.; Yang, M.; Pan, H. Aprender a mejorar Eficiencia operativa desde Pose Estimación de errores en robótica Polinización. Electrónica 2024, 13, 3070. https://doi.org/10.3390/electrónica13153070

Editor académico: Dah-Jye Lee

Recibido: 24 de junio de 2024
Revisado: 26 de julio de 2024
Aceptado: 1 de agosto de 2024
Publicado: 2 de agosto de 2024

1. Introducción

La polinización asistida por robots ha sido un tema durante más de una década [1]. Se está convirtiendo en un contribuyente cada vez más importante a la producción de cultivos en la industria agrícola y está brindando soluciones a desafíos apremiantes, como la disminución de los polinizadores naturales. Los robots de polinización se clasifican en varios tipos para satisfacer diversas necesidades agrícolas. Robots autónomos, equipados con visión avanzada e inteligencia artificial, navegan y polinizan de forma independiente. Los robots semiautónomos o controlados remotamente requieren supervisión humana, adecuados para entornos complejos. Los vehículos aéreos no tripulados (UAV) polinizan eficientemente grandes campos desde arriba. Estos sistemas robóticos están diseñados para imitar el proceso de polinización que tradicionalmente realizan las abejas y otros insectos. Utilizando sensores avanzados, sistemas de visión y manipuladores de precisión, los polinizadores robóticos identifican e interactúan con las flores, depositando polen con gran precisión y consistencia.

Se han discutido varios estudios en el campo de la polinización automatizada con aplicaciones en la polinización de kiwi [2,3], flores de zarza [4], vainilla [5] y flores de tomate [6]. La técnica de polinización requiere la transferencia física de polen del macho a los pistilos (los órganos reproductores de las flores femeninas) [7]. Para minimizar la pérdida de polen y asegurar una transferencia precisa a los estigmas, es necesario abordar las dificultades

de recolectar polen y los desafíos de mantener su actividad [8]. Desafortunadamente, esto requiere mucho tiempo y esfuerzo, lo cual no se ha considerado cuidadosamente en los enfoques de polinización existentes. Las dificultades provienen del tamaño muy pequeño de los pistilos, así como del desafío de identificar con precisión su rotación y orientación.

Para ello, este trabajo propone una metodología basada en una arquitectura transformadora que tiene como objetivo mejorar la eficiencia operativa de la polinización robótica en tareas de cruzamiento. La capa de salida del modelo incorpora dos redes neuronales de avance separadas, cada una diseñada para hacer una regresión de las discrepancias de traslación y rotación entre el robot y el objetivo de polinización, respectivamente. A diferencia de las funciones de pérdida de transformador tradicionales, hemos diseñado una función de pérdida novedosa que se basa en las discrepancias de traslación y rotación entre el efector final del robot y los objetivos de polinización. Esto permite que el brazo robótico identifique de forma rápida y precisa el objetivo de polinización desde la vista de observación actual. Mientras tanto, también hemos diseñado una estrategia para recopilar muestras de entrenamiento a gran escala a partir de la observación ojo en mano y designar los errores de traslación y rotación identificados dentro del sistema de coordenadas cartesiano del manipulador final como objetivos de pérdida. Las imágenes observadas en las posiciones actual y objetivo, combinadas con los errores de traslación y rotación entre ellas, proporcionan una gran escala para los datos de entrenamiento, mejorando así el entrenamiento del modelo. La Figura 1 muestra el robot de polinización utilizado en este trabajo.

El resto de este trabajo está organizado de la siguiente manera: los trabajos relacionados se describen en la Sección 2; las descripciones detalladas del método se presentan en la Sección 3; Las secciones 4 a 6 presentan los experimentos, la discusión y la conclusión, respectivamente.

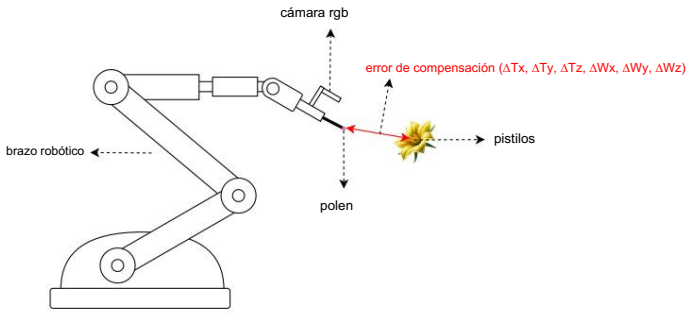


Figura 1. El robot de polinización consta principalmente de un brazo robótico UR5 y una cámara RGB monocular, configurados en una estructura de ojo en mano. Al final del brazo robótico, hay un cepillo de polinización que se utiliza para polinizar los pistilos. La ilustración muestra un brazo robótico dedicado a la polinización, donde el "error de compensación" marcado en rojo representa el contenido que el modelo necesita aprender. Aquí, ΔTx, ΔTy y ΔTz denotan los componentes del error de traslación a lo largo de los ejes X, Y y Z , respectivamente, mientras que ΔWx, ΔWy y ΔWz representan los componentes del vector rotacional, lo que indica la errores de rotación.

2. Trabajo relacionado

2.1. Estimación de postura

Lograr una alta precisión en la postura del efector final de un brazo robótico plantea desafíos importantes. Para mejorar la precisión de la posición del efector final, a menudo se emplean sensores adicionales, como LiDAR o sensores de presión, lo que puede aumentar los costos. Algunos métodos se basan en la estimación de la pose del objeto 6D, la construcción de plantillas para escanear diferentes posiciones en la imagen de entrada, el cálculo de puntuaciones de similitud en cada posición y la obtención de la mejor coincidencia comparando las puntuaciones de similitud [9,10]. Otros utilizan métodos basados en características, donde los modelos CNN hacen una regresión directa de las coordenadas 3D de cada píxel [11], o describen la densidad posterior de una pose de objeto específica a través de CNN, comparando la imagen observada con la imagen renderizada [12]. Además, algunos métodos combinan las ventajas de los enfoques basados en plantillas y en funciones dentro de un marco de aprendizaje profundo. La red integra el etiquetado de nivel de píxel de abajo hacia arriba con la regresión de pose de objeto de arriba hacia abajo, prediciendo

la posición 3D y la rotación 3D del objeto de forma desacoplada [13]. Todos estos métodos utilizan invariablemente sensores de profundidad para capturar la información de profundidad D del objeto objetivo. Sin embargo , en escenarios donde no se puede capturar información de profundidad, como en el robot de polinización agrícola analizado en este artículo, las regiones huecas de las flores de polinización dan como resultado una adquisición deficiente de información de profundidad. Además, los modelos entrenados que incorporan información de profundidad [11-13] no retroceden ni predicen directamente la información de posición 3D del objeto objetivo. Los modelos basados en plantillas y en características antes mencionados se entrenan en nubes de puntos, pero aún enfrentan desafíos al manejar poliedros no convexos y datos no múltiples. Los métodos de entrenamiento de nubes de puntos existentes son principalmente adecuados para formas regulares, como poliedros convexos y variedades. Estos métodos son limitados cuando se trata de formas complejas e irregulares de flores en ambientes agrícolas, que a menudo presentan estructuras no poliédricas y no múltiples. Esta limitación dificulta la aplicación directa de los métodos tradicionales de formación de nubes de puntos. Además, existe una discrepancia entre el entrenamiento de nubes de puntos en entornos virtuales y aplicaciones del mundo real, lo que puede afectar el rendimiento del modelo en escenarios del mundo real.

A diferencia de las técnicas de estimación de pose introducidas anteriormente, que dependen de sensores adicionales para capturar datos de imágenes con profundidad mejorada para construir un entrenamiento de nubes de puntos (una práctica difícil de implementar en entornos agrícolas no convexos y no múltiples), nuestro método utiliza sólo cámaras convencionales, para adquirir imágenes RGB del efector final del robot de polinización. Este enfoque permite una capacitación integral y fluida en entornos del mundo real, eliminando así las discrepancias que se encuentran comúnmente entre la capacitación en nubes de puntos en entornos virtuales y su aplicación práctica en el campo.

2.2. Métodos tradicionales y de aprendizaje profundo para mejorar la precisión

Para mejorar la precisión del efector final de polinización, se pueden implementar métodos tradicionales mediante el diseño del brazo robótico. Por ejemplo, LI K diseñó un brazo robótico para la polinización de kiwis [14], logrando una alta precisión al expresar la velocidad del efector final y las velocidades angulares conjuntas a través de la matriz jacobiana. Durante la planificación de trayectorias, la interpolación polinómica quíntica asegura la continuidad y suavidad de las curvas de velocidad y aceleración de cada articulación. Para el análisis de simulación se utilizan cinemática directa y métodos de Monte Carlo , para cubrir toda el área de polinización con el efector final. En el sistema de visión, se utiliza una cámara binocular [15] para obtener las coordenadas 3D de las flores en tiempo real, combinándola con el sistema de control para la planificación de trayectorias y el cálculo de puntos de interpolación, para controlar con precisión cada motor de articulación y lograr punto a punto. polinización precisa . Sin embargo, este brazo robótico se basa en cámaras binoculares para obtener las coordenadas 3D de las flores, y la luz intensa o las oclusiones pueden afectar la precisión del sistema

Los algoritmos tradicionales todavía pueden lograr ciertos efectos en dominios específicos de polinización agrícola. Un estudio reciente realizado por N Duc Tai [16] implicó el uso de métodos de segmentación para localizar y segmentar flores de melón, empleando modelos matemáticos basados en las características biológicas de las flores de melón para determinar los puntos clave de la dirección de crecimiento de cada flor y usando una proyección inversa. Método para convertir la posición de la flor de una imagen 2D a un espacio 3D, logrando así la localización de la pose de la flor.

En los últimos años, con el desarrollo y la aplicación del aprendizaje profundo, el robot de polinización BrambleBee [4] de STRADER J ha combinado algoritmos tradicionales y algoritmos de aprendizaje profundo para estimar las poses de las flores. El sistema utiliza un marco de procesamiento de imágenes de dos etapas para reconocer flores y estimar sus poses. Primero emplea un algoritmo de segmentación a nivel de píxeles de Naive Bayes, luego utiliza una red neuronal convolucional para la clasificación y el refinamiento de la pose, lo que garantiza la precisión de los resultados del reconocimiento. Finalmente, un mapa de obstáculos basado en octree y un gráfico de factores representan el mapa de flores, mapeando las flores y sus posiciones en 3D para optimizar la estimación de la pose. El robot de polinización automática de flores de sandía de Khubaib Ahmad [17] utiliza el aprendizaje profundo para detectar la información 2D de las flores y la combina con algoritmos tradicionales para calcular la profundidad de las posiciones de las flores, logrando así la localización de las flores. Luego, el robot ajusta el brazo mecánico, mediante servobúsqueda, hasta que alcanza la posición

El robot de polinización servoautomatizado basado en visión de YANG [18] utiliza algoritmos de detección de objetivos más avanzados, como YOLOv5, YOLACT++ [19] y DETR [20], para detectar flores, identificando la posición y orientación del pistilo mediante la rotación. tecnología de detección de objetos . Luego, el sistema emplea una estrategia de alcance pseudobinocular, calculando las coordenadas 3D del pistilo moviendo la posición de la cámara y usando la calibración ojo-mano para transformar estas coordenadas en el sistema de coordenadas operativas del brazo robótico . Durante la tarea de polinización, el sistema combina estrategias de servocontrol visual, realizando un posicionamiento aproximado moviendo el efector final cerca de la flor, seguido de un posicionamiento preciso utilizando una trayectoria de búsqueda circular para garantizar un contacto preciso entre el cepillo de polen y el pistilo. Sin embargo, el sistema todavía tiene un error de precisión del efector final de 15 mm. Aunque la trayectoria de búsqueda circular puede eventualmente lograr un posicionamiento preciso, el proceso lleva mucho tiempo y tarda casi 19 s en completar la polinización de una sola flor.

Si se puede mejorar la precisión del efector final, la eficiencia del posicionamiento preciso a través de la trayectoria de búsqueda circular se beneficiará significativamente.

Tanto los métodos tradicionales de N Duc Tai como los robots de polinización basados en aprendizaje profundo de STRADER J, Khubaib Ahmad y YANG impactan significativamente la tasa de éxito y la eficiencia de las operaciones de polinización a través de errores en la estimación de la postura de los objetivos de polinización . Por ejemplo, los robots de polinización de N Duc Tai y YANG solo pueden garantizar una polinización exitosa manteniendo un área de búsqueda servo grande, lo que conduce a una eficiencia de polinización reducida. En consecuencia, nuestro trabajo propone un método de aprendizaje profundo para aprender los errores de traslación y rotación entre la flor y el efector final de la polinización. Al compensar estos errores, se mejora la precisión de posicionamiento del efector final del robot. Esta reducción en el rango de búsqueda y la ruta del servomecanismo en última instancia mejora la eficiencia del proceso de polinización.

2.3. Arquitectura de transformador y codificación de posición

La arquitectura transformadora, introducida por Vaswani et al. [21], ha tenido un impacto significativo en el procesamiento del lenguaje natural (PLN) y la visión por computadora. Los transformadores utilizan el mecanismo de autoatención para capturar dependencias complejas dentro de los datos de entrada, haciéndolos efectivos para comprender la información visual. En visión por computadora, los transformadores procesan imágenes segmentándolas en parches e incrustándolas linealmente, lo que permite poderosos mecanismos de atención para tareas como clasificación de imágenes, detección de objetos y generación de descripciones de imágenes. En robótica, los transformadores pueden predecir las compensaciones de error del extremo de un brazo robótico analizando secuencias de imágenes o datos de sensores, mejorando la precisión y la exactitud.

La introducción de la codificación de posición [22] para predecir el error de compensación de traslación y el error de compensación de postura de rotación en el extremo del brazo robótico mejora aún más la conciencia espacial y la precisión. Al incorporar información de posición espacial directamente en el modelo del transformador, se mejora la capacidad del modelo para discernir las posiciones relativas y absolutas de los objetos en una escena, lo cual es fundamental para una estimación precisa del error.

El mecanismo de autoatención de los transformadores puede capturar dependencias de largo alcance, lo que permite que el modelo calcule directamente las relaciones entre dos elementos cualesquiera dentro de una secuencia de características de la imagen de entrada, considerando todos los datos de entrada en lugar de solo la información local . Esta característica ayuda a reducir el problema de los óptimos locales. Además, mediante una asignación de atención detallada, los modelos basados en transformadores pueden identificar y enfatizar las características y relaciones más relevantes para la tarea actual. Por lo tanto, emplear un modelo de transformador equipado con codificación posicional para predecir los errores de traslación y rotación entre el efector final del robot de polinización y el objeto objetivo es un enfoque viable.

3. Método

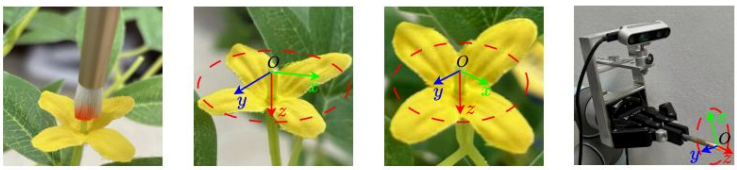
3.1. Error de

compensación Para describir el error de compensación de traslación y el error de compensación de pose rotacional, se deben construir dos sistemas de coordenadas cartesianas, CA y CB, en la flor objetivo y en

el efector final del robot de polinización, respectivamente. El sistema de coordenadas cartesiano CA, construido sobre la flor, tiene su origen O en el extremo del pistilo. El plano formado por los ejes xey es paralelo al plano de los pétalos de la flor, y el eje z es paralelo al pistilo, apuntando hacia adentro, como se muestra en la Figura 2b,c. El sistema de coordenadas cartesianas CB, construido al final del robot de polinización, tiene su origen O al final del cepillo, obtenido al traducir el sistema de coordenadas TCP del brazo robótico UR5, como se muestra en la Figura 2d. CA se puede derivar de CB a través de una matriz de traducción-rotación TR de 4 × 4:

$$TR = \begin{matrix} R11 & R12 & R13 & t1 \\ R21 & R22 & R23 & t2 \\ R31 & R32 & R33 & t3 \\ 0 & 0 & 0 & 1 \end{matrix} \quad (1)$$

$$CA = TR \cdot CB \quad (2)$$



(a) (b) (C) (d)

Figura 2. (a) La posición ideal de polinización sin error. (b – d) Las posiciones de los sistemas de coordenadas cartesianas en la flor y al final del brazo robótico, respectivamente.

El error de desplazamiento de traslación y el error de desplazamiento de pose de rotación son el vector de traslación t y el vector de rotación derivado de la matriz de rotación R de la matriz de traslación-rotación TR, respectivamente:

$$t = t1\ t2\ t3 \quad (3)$$

$$R = \begin{matrix} R11 & R12 & R13 \\ R21 & R22 & R23 \\ R31 & R32 & R33 \end{matrix} \quad (4)$$

Cuando el efector final del robot de polinización está en la postura de polinización ideal, los valores del error de compensación de traslación y el error de compensación de postura de rotación se acercan a cero. En este momento, el cepillo al final del robot de polinización está perpendicular al plano de los pétalos de la flor objetivo y justo en contacto con el pistilo, es decir, los sistemas de coordenadas CA y CB coinciden, como se muestra en la Figura 2a.

$$\theta = \cos^{-1}\left(\frac{\text{traza}(R) - 1}{2}\right)$$

$$\vec{V} = \frac{1}{2\,\text{pecado}(\theta)} \begin{matrix} R32 - R23 \\ R13 - R13 \\ R21 - R12 \end{matrix} \quad (5)$$

$$\vec{W} = \theta\ \vec{V}$$

El error de compensación de traslación y el error de compensación de pose rotacional predichos por el modelo se denotan como $\hat{t}$ y $\hat{R}$, respectivamente. Las diferencias entre el error de compensación de traslación real y previsto y el error de compensación de postura de rotación desde el extremo del brazo robótico hasta el pistilo de la flor objetivo se calculan como el error de traslación (TE) y el error de rotación (RE), respectivamente. Según la ecuación (5), las matrices de rotación R y $\hat{R}$ se pueden convertir en vectores de rotación $\vec{W}$ y $\hat{\vec{W}}$:

$$TE = \hat{t} - t \quad (6)$$

$$RE = 1 - \frac{\vec{W}}{\vec{W}\ \vec{W}\ \vec{W}} \quad (7)$$

La unidad de TE es centímetros (cm), que representa la distancia espacial entre el error de traducción previsto y el error de traducción real. RE no tiene dimensiones y representa la distancia del coseno entre el error del vector de rotación predicho y el error del vector de rotación real.

3.2. Módulo de Atención

El módulo de atención se implementa sobre la base del mecanismo de autoatención, que permite al modelo sopesar la importancia de diferentes características dentro de los datos de entrada.
Al aplicar este mecanismo, el modelo puede concentrar de forma adaptativa recursos computacionales en regiones de la imagen que contienen información clave para predecir compensaciones de pose (ΔTx, ΔTy, ΔTz, ΔWx, ΔWy, ΔWz). Además, con la introducción de la codificación de posición en el mecanismo de autoatención, el modelo logra una comprensión integral de la imagen. Esto proporciona información adicional para abordar problemas de simetría en la predicción de la pose de los objetos, mejorar la precisión del modelo y aumentar su capacidad general en diversos escenarios encontrados por los robots de polinización. Este método no sólo mejora la precisión del modelo sino que también fortalece su versatilidad en las diversas condiciones que enfrentan los robots de polinización.

3.3. Extracción de características

Dada la capacidad única de los modelos de transformadores para procesar datos de imágenes, utilizamos una red neuronal convolucional previamente entrenada, ResNet-50 [23], como red de extracción de características, para convertir imágenes originales en vectores de características de alta dimensión. Estas características luego se introducen en el modelo del transformador para su posterior procesamiento. Comparamos redes de extracción de características de diferentes profundidades para determinar la representación óptima de las características. Este enfoque aprovecha la gran capacidad de ResNet-50 para capturar jerarquías espaciales detalladas en imágenes, proporcionando un rico conjunto de características para que el modelo de transformador las analice. Al incorporar esta arquitectura híbrida, combinando las fortalezas de las CNN en la extracción de características con el mecanismo de atención avanzado de los transformadores, el modelo logra una comprensión matizada del contenido de la imagen relevante para predecir los errores de compensación de pose. Este método facilita la identificación y el enfoque en aspectos cruciales de los datos de entrada, mejorando así la precisión y eficiencia del proceso de estimación de pose.

3.4. Enfoque propuesto

Este trabajo presenta un modelo de red novedoso que incorpora un módulo de atención, diseñado para mejorar la precisión de la predicción centrándose en las partes más críticas de los datos de entrada. Este modelo es particularmente adecuado para analizar datos de imágenes capturados por el efector final de un robot de polinización, con el objetivo de predecir con precisión el desplazamiento de posición del efector final en relación con el objeto objetivo en un sistema de coordenadas cartesianas. Durante el entrenamiento, el modelo toma datos de imagen del extremo del brazo robótico como entrada y genera un vector de seis dimensiones (ΔTx, ΔTy, ΔTz, ΔWx, ΔWy, ΔWz) que representa el desplazamiento traslacional $\vec{W}$ = (ΔWx, ΔWy, ΔWz) en los errores ΔT = (ΔTx, ΔTy, ΔTz) y los errores de desplazamiento rotacional en tres direcciones del sistema de coordenadas cartesiano.

3.5. Predicción de error de

compensación La entrada al modelo es una imagen RGB capturada por una cámara RGB "ojo en mano" en el brazo robótico, que representa el estado posicional del efector final del brazo robótico durante la polinización y la flor de polinización. La salida es un vector que representa el vector de desplazamiento de pose (ΔTx, ΔTy, ΔTz, ΔWx, ΔWy, ΔWz), que incluye los errores de traslación y rotación del efector final del robot en relación con el objeto objetivo. Para lograr esto, el modelo emplea dos redes neuronales de retroalimentación para mapear directamente las características centradas en la atención.

en el espacio de desplazamiento tridimensional y el espacio de rotación tridimensional. La Figura 3 ilustra todo el proceso algorítmico.

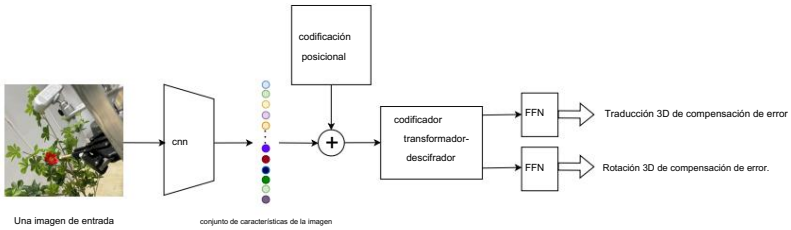


Figura 3. El diagrama ilustra el flujo de trabajo del algoritmo propuesto. El proceso comienza con la imagen de entrada, que se somete a extracción de características a través de una red neuronal convolucional (CNN), para obtener características de imagen de alta dimensión. Luego, estas características se aumentan con codificación posicional antes de ingresarlas al módulo transformador. La salida del transformador es procesada adicionalmente por dos redes neuronales de avance distintas, encargadas de predecir errores de traslación y rotación, respectivamente.

3.6. Red de arquitectura

En este trabajo, hemos adoptado un modelo de transformador personalizado diseñado para procesar características de imágenes serializadas y predecir el error de compensación de traslación y el error de compensación de pose rotacional en el extremo de un brazo robótico. Hemos modificado el modelo de transformador original introduciendo codificación posicional bidimensional, para preservar la información espacial de las imágenes de entrada. La capa de salida está personalizada para generar un desplazamiento de traslación 3D y un vector de error de desplazamiento de pose rotacional. Como se muestra en la Figura 4, un ResNet50 sirve como columna vertebral para extraer un rico conjunto de características de los datos de la imagen de entrada. Para retener la información espacial entre los elementos de las imágenes, aplicamos codificación de posición sinusoidal a las características extraídas, que luego se suman con las características antes de int
Para capturar información de diferentes subespacios, en el codificador se emplea un mecanismo de atención de múltiples cabezales con cinco cabezales. La salida del codificador se envía posteriormente al decodificador. Siguiendo la arquitectura estándar del transformador, el decodificador utiliza un mecanismo de atención de múltiples cabezales para transformar dos incrustaciones de tamaño d. El decodificador transforma estos objetos de consulta en incrustaciones de salida, que luego son decodificadas por dos redes de avance separadas en el error de compensación de traslación y el error de compensación de pose rotacional, respectivamente.

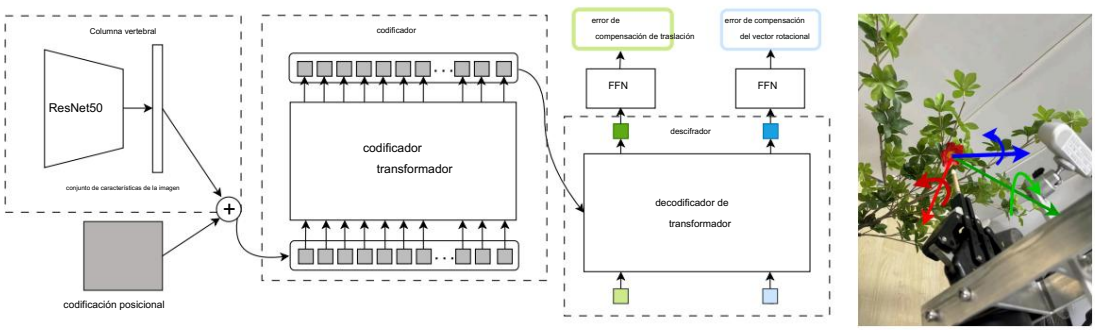


Figura 4. El modelo final emplea ResNet50 como red troncal para aprender características de alto nivel a partir de la imagen de entrada. Luego, estas características se complementan con codificación posicional antes de pasarlas al codificador. Posteriormente, el decodificador genera primero un vector de características para el error de compensación de traducción, que se utiliza como entrada para una red neuronal de avance diseñada para la predicción. Luego, este vector de características se utiliza como entrada de consulta para el decodificador, lo que da como resultado otro vector de características que se introduce en una red neuronal de alimentación directa separada para predecir el error de compensación de la post

4. Experimentos

4.1. Preprocesamiento de datos

Los datos de imagen utilizados en este experimento provienen todos de un conjunto de datos patentado, que Consiste en imágenes capturadas por una cámara montada en el brazo robótico durante la polinización. proceso, donde cada imagen representa de forma única la pose de una flor que será polinizada. Como se muestra en la Figura 5, la orientación de las flores con respecto al extremo del brazo robótico Se clasificó en cinco direcciones: izquierda, derecha, arriba, abajo y frente. Detallado La información estadística está disponible en la Tabla 1.

(Hacia arriba) (hacia abajo) (Frente) (Izquierda) (Bien)

Figura 5. Flores con diferentes orientaciones con respecto al efector final del brazo robótico.

Tabla 1. El conjunto de datos incluye el número de imágenes de flores orientadas en diferentes direcciones en relación al efector final del brazo robótico.

Orientaciones	Número de flores	Proporción (%)
I	261	20.6
R	274	21.7
F	230	18.2
Ud.	256	20.3
D	243	19.2

"F", "L", "R", "U" y "D" representan las orientaciones frontal, izquierda, derecha, hacia arriba y hacia abajo.

Para garantizar la coherencia de los datos de entrada, todas las imágenes capturadas por el brazo robótico La cámara final primero se redimensionó a una resolución uniforme de (224 × 224). Posteriormente, las imágenes se normalizaron, escalando los valores de píxeles al rango [0, 1] , para mejorar la estabilidad del entrenamiento del modelo. Además, se aplicó una serie de técnicas de aumento de datos, que incluyen escalado, recorte y transformación de color, para aumentar la diversidad de datos y evitar el sobreajuste. En cuanto a los datos de la etiqueta, asumimos la establecer el error $\Delta T = (\Delta T_x, \Delta T_y, \Delta T_z)$ y el error de compensación rotacional $\Delta \vec{W} = (\Delta W_x, \Delta W_y, \Delta W_z)$, con $||\Delta T|| < D$, donde D era una constante. El error de compensación de posición Ecuación (8) y El error de compensación rotacional de la ecuación (9) se normalizó y se escaló a [−1, 1], de la siguiente manera:

$$\Delta T_n = \frac{\Delta T}{D} \tag{8}$$

$$\vec{W} = \theta \vec{V} \tag{9}$$

$$\theta = \sqrt{\Delta W_x^2 + \Delta W_y^2 + \Delta W_z^2} \tag{10}$$

$$\vec{V} = \left(\frac{\Delta W_x}{\theta}, \frac{\Delta W_y}{\theta}, \frac{\Delta W_z}{\theta}\right) \tag{11}$$

El símbolo θ en la ecuación (10) representa el ángulo de rotación, y $\vec{V}$ denota el vector unitario a lo largo del eje de rotación. Los datos de etiqueta normalizados fueron $(\Delta T_n, \vec{V}, \frac{\theta}{2\pi})$.

4.2. Métricas de evaluación

En el modelo presentado en este artículo, predecimos por separado el desplazamiento traslacional error y el error de compensación de pose rotacional, diseñando así dos funciones de pérdida. la primera derrota

La función, denominada LossT, mide la distancia media cuadrática entre el error de compensación de la posición espacial predicho por el modelo para el efector final del brazo robótico de polinización hasta el pistilo de la flor objetivo y el error de compensación espacial real. LossT se define de la siguiente manera:

$$\text{PérdidaT} = \frac{1}{2mx} \sum_{M} (\hat{T}_{norte} - Tn)^2 \quad (12)$$

donde M es el conjunto del conjunto de datos de prueba, m es el número de elementos en el conjunto y $\hat{T}$ Tn son $\hat{T}_{norte}$ y el error de compensación de traslación predicho por el modelo y el error de compensación de traslación verdadero obtenido mediante la Ecuación (8), respectivamente.

La segunda función de pérdida, denominada LossR, mide la discrepancia entre el error de rotación espacial predicho por el modelo para el efector final del brazo robótico de polinización del pistilo de la flor objetivo y el error de rotación espacial real. LossR se define de la siguiente manera:

$$\text{PérdidaR} = \frac{1}{2m_x} \sum_{M} \left(\frac{1}{\sigma_1^2} \left(\hat{\theta} - \theta \right)^2 \right) + \frac{1}{\sigma_2^2} \left(1 - \frac{\vec{V} \cdot \hat{\vec{V}}}{\vec{V} \quad \hat{\vec{V}}}\right) + \log(\sigma_1 \sigma_2)) \quad (13)$$

donde M es el conjunto del conjunto de datos de prueba y m es el número de elementos del conjunto. Variables $\hat{\theta}$, θ, $\hat{\vec{V}}$, y $\hat{\vec{V}}$ representan los ángulos y ejes de rotación previstos y reales obtenidos mediante Ecuaciones (9)–(11), siendo σ 1 y σ 2 los parámetros que deben aprenderse.

La función de pérdida combinada utilizada para el entrenamiento del modelo se define como

$$\text{Pérdida} = \alpha\text{PérdidaT} + \beta\text{PérdidaR} \quad (14)$$

Esta función de pérdida en la Ecuación (14) mide de manera integral la pérdida por error de compensación traslacional y la pérdida por error de compensación rotacional durante el entrenamiento del modelo. Los parámetros α y β son hiperparámetros que representan pesos que deben ajustarse sistemáticamente durante el entrenamiento del modelo.

4.3. Detalles de la

capacitación La capacitación del modelo se llevó a cabo en un entorno computacional equipado con GPU NVIDIA V100. Usamos el optimizador Adam [24], con la tasa de aprendizaje inicial establecida en 0,01, y empleamos una estrategia de disminución de la tasa de aprendizaje que redujo gradualmente la tasa de aprendizaje a 0,00001 a medida que avanzaba el entrenamiento. Durante el proceso de capacitación se utilizó una función de pérdida personalizada, la ecuación (14) . En el entrenamiento del modelo, la configuración de los hiperparámetros en la Ecuación (14) afectó la capacidad del modelo para converger. Después de extensos experimentos, finalmente establecimos α = 0,0025 y β = 1, lo que permitió que el modelo convergiera más fácilmente durante el entrenamiento.

En el conjunto de datos, se mezclaron aleatoriamente flores con diferentes orientaciones y luego los datos se dividieron en 10 subconjuntos, utilizando el método de validación cruzada K-fold. Cada vez, se utilizó un subconjunto como conjunto de prueba y los nueve subconjuntos restantes como conjunto de entrenamiento. Este proceso se repitió 10 veces. Durante el entrenamiento del modelo, observamos que el modelo inicialmente convergió rápidamente y luego se estabilizó gradualmente. La Figura 6 muestra los cambios en los valores de pérdida, error de traslación y error de rotación durante el proceso de entrenamiento del modelo:

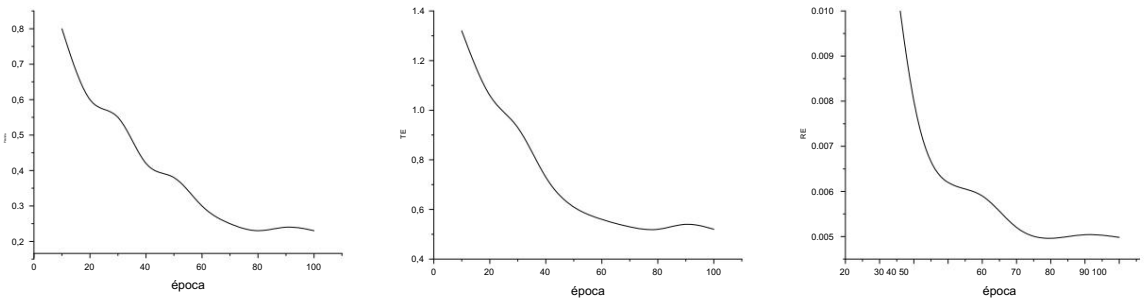


Figura 6. Cambios en pérdida, TE y RE durante el proceso de entrenamiento del modelo.

4.4. Resultados

Hasta donde sabemos, somos los primeros en predecir el error de compensación de traslación. y error de compensación de postura rotacional del efector final de un brazo robótico en relación con un objeto objetivo usando Sólo información de imagen RGB. Por lo tanto, este artículo se centra en las mejoras de precisión. del error de compensación de traslación y del error de compensación de pose rotacional del efector final del robot cuando utilizamos nuestro método propuesto en comparación con el método YANG de última generación [18] como la línea de base. Además, destacamos la mejora en la eficiencia de polinización de un sola flor.

Realizamos experimentos con flores con diferentes orientaciones en grupos, para analizar errores de compensación de traslación y rotación, así como la velocidad de detección. Los resultados experimentales mostró ligeras variaciones para flores con diferentes orientaciones. Como se muestra en la Tabla 2 y Figura 7, las flores orientadas hacia adelante lograron los mejores resultados, en términos de precisión experimental, con el menor error de compensación de traslación y error de compensación de postura de rotación en comparación con flores en otras orientaciones. Luego vinieron las flores mirando hacia arriba. Debido al medio ambiente simetría, casi no hubo diferencia en los resultados para las flores orientadas hacia la izquierda y hacia la derecha. flores hacia abajo tuvo los peores resultados, tanto en términos de error de compensación de traslación como error de compensación de pose rotacional. Sin embargo, la velocidad de detección del modelo para flores con las diferentes orientaciones permanecieron casi constantes.

Tabla 2. Resultados experimentales del modelo sobre flores con diferentes orientaciones.

Orientaciones	TE	RE	FPS
I	0,82	0,0049	41
R	0,82	0,0049	42
F	0,78	0,0047	43
Ud.	0,80	0,0048	42
D	0,87	0,0051	41

“F”, “L”, “R”, “U” y “D” representan las orientaciones frontal, izquierda, derecha, hacia arriba y hacia abajo.

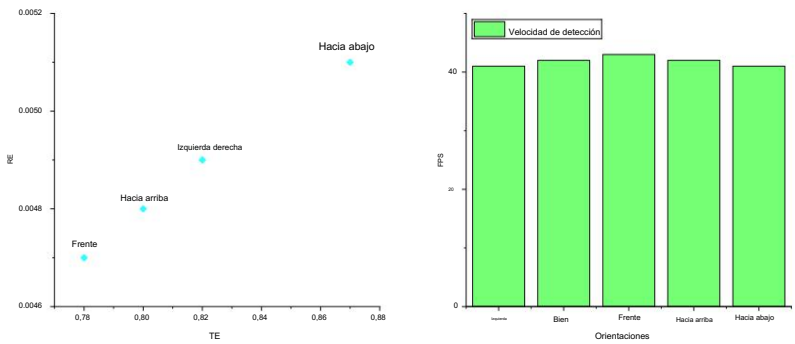


Figura 7. Distribución de la detección de errores rotacionales y traslacionales del modelo para flores con diferentes orientaciones, así como la velocidad de detección.

Realizamos nuestros experimentos utilizando el robot de polinización YANG. El proceso de polinización de una sola flor del robot de polinización YANG se puede dividir en cinco pasos en

orden cronológico. Después del cuarto paso, la precisión de posicionamiento del efector final del robot de polinización YANG en relación con el pistilo objetivo fue de 1,5 cm. En este punto, presentamos un nuevo paso llamado “Posición de ajuste fino”. Ingresamos la imagen de la pose de la flor. en relación con el efector final del robot en nuestro modelo entrenado, para predecir la traslación y Errores de compensación rotacional. Según los valores previstos, la pose del robot de polinización Se ajustó el efector final.

Como se muestra en la Tabla 3, gracias al nuevo paso "Posición de ajuste fino", la distancia traslacional y la discrepancia rotacional entre el efector final del robot de polinización y el objetivo de polinización se redujo aún más, reduciendo el rango para la siguiente búsqueda de servo paso de polinización. Nuestros experimentos mostraron que el tiempo promedio para la polinización por búsqueda servo fue de sólo 3,1 s, logrando una tasa de éxito de polinización comparable a la polinización de YANG. robot al 86,19%. Como se muestra en la Tabla 4, después de aplicar nuestro método al robot de polinización, la precisión de distancia promedio entre el efector final del robot y el pistilo objetivo alcanzado 0,81 cm, una mejora del 46,67%. El error de compensación de la postura rotacional se calcula de acuerdo con a la Ecuación (14) también alcanzó 0,0049. El tiempo total para completar la polinización de un solo flor se redujo casi a la mitad, con una mejora promedio de la eficiencia del 50,9%.

Tabla 3. Comparación del costo de tiempo promedio para cada paso del sistema de polinización después de incorporar nuestro método.

Paso	YANG (línea de base) Costo de tiempo (S)	Nuestro costo de tiempo (S)
Detección de flores	0.0928	0.0927
Identificación del pistilo	0.025	0.024
Cálculo de posición (movimiento del robot incluido)	1.8	1.8
Alcance de flores (movimiento de robot incluido)	4.2	4.2
Posición de ajuste fino	/	0.024
Servo (movimiento del robot incluido)	12.7	3.1

El paso “Posición de ajuste fino” es un paso adicional. Con este paso incluido, el tiempo empleado en el “Servoing” El paso se reduce considerablemente, logrando la misma tasa de éxito de polinización en solo 3,1 s. Todos los números están registrados. en segundos.

Tabla 4. Comparación del error de compensación de traslación y el error de compensación de pose rotacional al aplicar nuestra algoritmo para flores orientadas en diferentes direcciones en comparación con la línea de base.

Orientaciones	Método	TE	RE	Tiempo Costo (s)
I	YANG	1,53	/	18,78
	Nuestro	0,82	0,0050	9.22
R	YANG	1,52	/	18.80
	Nuestro	0,82	0,0050	9.23
F	YANG	1,48	/	18,85
	Nuestro	0,80	0,0048	9.22
Ud.	YANG	1,53	/	18,79
	Nuestro	0,81	0,0049	9.24
D	YANG	1,55	/	18.83
	Nuestro	0,83	0,0052	9.26

“F”, “L”, “R”, “U” y “D” representan las orientaciones frontal, izquierda, derecha, hacia arriba y hacia abajo. El método de Yang sirvió como base.

Este artículo realizó varios experimentos de ablación para verificar el impacto de diferentes backbones y la adición de codificación posicional en la predicción del modelo del error de compensación de traslación y el error de compensación de pose rotacional. ResNet50, ResNet18, ResNet101, Se seleccionaron VGG16, VGG19, DenseNet-121 y DenseNet-201 como función principal Redes de extracción. Para cada columna vertebral, se realizaron experimentos con y sin codificación posicional. Los resultados experimentales, que se muestran en la Tabla 5 y la Figura 8, indican que las diferentes redes troncales y la adición de codificación posicional afectan significativamente la

precisión de la predicción final del modelo. La versión del modelo con ResNet50 como columna vertebral.
La red de extracción de características y la codificación posicional agregada lograron el mejor rendimiento.
en la predicción del error de compensación de traslación y el error de compensación de pose rotacional, alcanzando
8,1 mm y 0,0049, respectivamente.

Tabla 5. El impacto de diferentes pilares dentro del modelo en la precisión de predecir el
error de compensación de traslación y error de compensación de pose rotacional.

Columnas vertebrales	Codificación posicional	TE	RE
ResNet50		0,81	0.0049
		1.19	0.0051
ResNet18		0,92	0.0053
		1.33	0.0054
ResNet101		0,86	0.0050
		1.21	0.0050
VGG16		0,96	0.0053
		1.32	0.0055
VGG19		0,99	0.0054
		1.33	0.0055
DensoNet-121		0,92	0.0054
		1.23	0.0055
DensoNet-201		1.10	0.0055
		1.38	0.0057

Una marca de verificación () indica que el modelo incluía codificación posicional, mientras que una cruz () indica que el
El modelo no incluía codificación posicional.

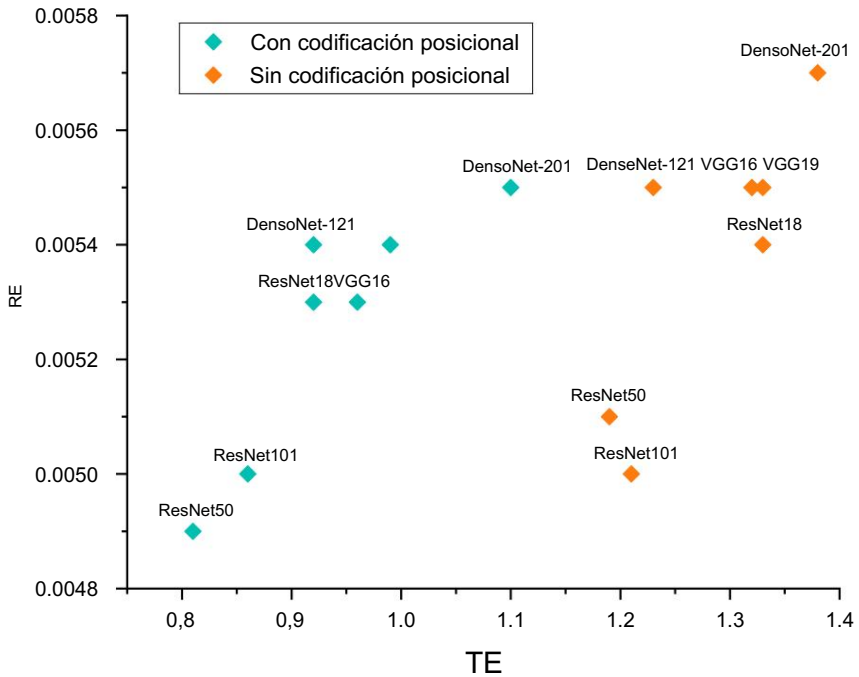


Figura 8. Distribución del error de traslación y rotación para diferentes modelos troncales con y
sin la adición de codificación posicional.

5. Discusión

Este trabajo propone un enfoque basado en transformadores que logra la predicción de un extremo a otro.
de errores de traslación y rotación entre el efector final del brazo robótico de polinización

y la posición de polinización objetivo, utilizando únicamente información de imagen RGB. Nuestros resultados experimentales demuestran que este método reduce efectivamente el rango de error del efector final robótico dentro de un margen de error conocido, mejorando así la eficiencia general de la polinización robótica.

En este trabajo, nuestro método mostró ligeras variaciones en los resultados al tratar con flores de diferentes orientaciones, particularmente errores de traslación y rotación más grandes con flores orientadas hacia abajo. Esto puede atribuirse al posicionamiento relativo del ángulo de la cámara del robot con respecto a la orientación de la flor, lo que complica el reconocimiento y la localización precisos de las flores en orientaciones específicas. A través de la experimentación, también observamos que las versiones del modelo que utilizaban codificación posicional tuvieron mejores resultados en la predicción precisa de errores de traslación y rotación. Además, las diferentes redes troncales utilizadas para extraer características de las imágenes de entrada tuvieron un impacto significativo en el rendimiento del modelo.

Este trabajo tiene limitaciones potenciales a pesar de la eficacia demostrada del método propuesto para reducir los errores de traslación y rotación entre el efector final del robot de polinización y el objetivo: (1) El conjunto de datos utilizado en este estudio se recopiló específicamente con fines experimentales, lo que puede limitar las capacidades de generalización del modelo en diversos ambientes y tipos de flores; (2) El rendimiento del modelo puede verse comprometido en diferentes condiciones de iluminación y en entornos obstruidos, lo que afecta su efectividad general. Estos problemas enfatizan la necesidad de realizar más mejoras antes de la implementación práctica, lo que requiere investigaciones futuras para explorar métodos adicionales que mejoren la adaptabilidad y confiabilidad del modelo en diversos entornos agrícolas.

6. Conclusiones

Este trabajo presenta un método innovador que utiliza las poderosas capacidades de comprensión y aprendizaje espacial de un modelo de aprendizaje profundo basado en transformadores para lograr una predicción de extremo a extremo de errores de traslación y rotación entre el efector final del robot de polinización y la posición de polinización objetivo utilizando solo RGB. imágenes. Nuestros resultados experimentales demuestran que este método es eficaz para reducir aún más el rango de error del efector final del robot de polinización dentro de un rango de error conocido, mejorando así la eficiencia general del robot de polinización. El trabajo futuro podría centrarse en investigar la predicción de errores de traslación y rotación entre el efector final robótico y las posiciones objetivo dentro de conjuntos de datos que contengan una gama más diversa de tipos de flores, bajo diferentes condiciones de iluminación y oclusiones. Dicho trabajo tendría como objetivo mejorar la solidez y las capacidades de generalización del modelo, proporcionando un enfoque factible para mejorar la precisión de los efectores finales robóticos genéricos.

Contribuciones de los autores: Conceptualización, JX; metodología, JX; validación, JX; análisis formal, JX, JC, HP y MY; investigación, JX; recursos, JX; curación de datos, JX; redacción: preparación del borrador original, JX, HP y MY; redacción: revisión y edición, JX, HP y MY; visualización, JX; supervisión, HP, MY y JC; administración del proyecto, JC Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: Este trabajo fue financiado por el Proyecto del Plan Clave de Investigación y Desarrollo de Guangxi (AB24010164; AB21220038).

Declaración de disponibilidad de datos: Los datos que respaldan los hallazgos de este trabajo están disponibles del autor correspondiente previa solicitud razonable.

Conflictos de intereses: Los autores declaran no tener conflictos de intereses.

Abreviaturas

En este manuscrito se utilizan las siguientes abreviaturas:

error de traducción TE
Error de rotación RE

Referencias

1. Binns, C. Los insectos robóticos podrían polinizar flores y encontrar víctimas de desastres. Ciencia popular, 17 de diciembre de 2009.
2. Williams, H.; Nejati, M.; Hussein, S.; Penhall, N.; Lim, JY; Jones, MH; Bell, J.; Ahn, HS; Bradley, S.; Schaare, P.; et al. Polinización autónoma de flores de kiwi individuales: hacia un polinizador robótico de kiwi. J. Robot de campo. 2020, 37, 246–262. [Referencia cruzada]
3. Gao, C.; Él, L.; Colmillo, W.; Wu, Z.; Jiang, H.; Li, R.; Fu, L. Un novedoso robot de polinización para flores de kiwi basado en preferencias Selección de flores y objetivo preciso. Computadora. Electrón. Agrícola. 2023, 207, 107762. [Referencia cruzada]
4. Strader, J.; Nguyen, J.; Tatsch, C.; Du, Y.; Lassak, K.; Buzzo, B.; Watson, R.; Cerbone, H.; Ohi, N.; Yang, C.; et al. Subsistema de interacción floral para un robot de polinización de precisión. En Actas de la Conferencia Internacional IEEE/RSJ de 2019 sobre Robots y Sistemas Inteligentes (IROS), Macao, China, 3 a 8 de noviembre de 2019; págs. 5534–5541.
5. Shaneyfelt, T.; Jamshidi, MM; Agaian, S. Un sistema de grúa de acoplamiento robótica con retroalimentación visual con aplicación a la polinización de la vainilla. En t. J. Autom. Controlar 2013, 7, 62–82. [Referencia cruzada]
6. Yuan, T.; Zhang, S.; Sheng, X.; Wang, D.; Gong, Y.; Li, W. Un robot de polinización autónomo para el tratamiento hormonal de flores de tomate en invernadero. En Actas de la Tercera Conferencia Internacional sobre Sistemas e Informática (ICSAI) de 2016, Shanghai, China, 19 a 21 de noviembre de 2016; págs. 108-113.
7. Abrol, DP Biología de la polinización: Conservación de la biodiversidad y producción agrícola: Volumen 792; Springer: Nueva York, NY, Estados Unidos, 2012.
8. Yang, X.; Miyako, E. Polinización con pompas de jabón. Iciencia 2020, 23, 101188. [CrossRef] [PubMed]
9. Hinterstoisser, S.; Cagniart, C.; Ilic, S.; Sturm, P.; Navab, N.; Fua, P.; Lepetit, V. Mapas de respuesta de gradiente para la detección en tiempo real de objetos sin textura. Traducción IEEE. Patrón Anal. Mach. Intel. 2011, 34, 876–888. [Referencia cruzada] [PubMed]
10. Cao, Z.; Jeque, Y.; Banerjee, NK Estimación de pose 6dof escalable en tiempo real para objetos sin textura. En Actas de la Conferencia Internacional IEEE sobre Robótica y Automatización (ICRA) de 2016, Estocolmo, Suecia, 16 a 21 de mayo de 2016; págs. 2441–2448.
11. Brachmann, E.; Krull, A.; Michel, F.; Gumhold, S.; Shotton, J.; Rother, C. Aprendizaje de la estimación de la pose de objetos 6D utilizando coordenadas de objetos 3D. En Actas, Parte II 13, Actas de Computer Vision – ECCV 2014: 13.ª Conferencia Europea, Zurich, Suiza, 6 a 12 de septiembre de 2014; Springer: Berlín/Heidelberg, Alemania, 2014; págs. 536–551.
12. Krull, A.; Brachmann, E.; Michel, F.; Yang, MI; Gumhold, S.; Rother, C. Análisis de aprendizaje por síntesis para la estimación de pose 6d en imágenes rgb-d. En Actas de la Conferencia Internacional IEEE sobre Visión por Computadora, Santiago, Chile, 7 a 13 de diciembre de 2015; págs. 954–962.
13. Xiang, Y.; Schmidt, T.; Narayanan, V.; Fox, D. Posecnn: Una red neuronal convolucional para la estimación de la pose de objetos 6d en escenas abarrotadas. arXiv 2017, arXiv:1711.00199.
14. Li, K.; Huo, Y.; Liu, Y.; Shi, Y.; Él, Z.; Cui, Y. Diseño de un brazo robótico liviano para la polinización de kiwis. Computadora. Electrón. Agrícola. 2022, 198, 107114. [Referencia cruzada]
15. Mamá, WP; Li, WX; Cao, PX Método de posicionamiento de objetos con visión binocular para robots basado en coincidencia estéreo gruesa-fina. En t. J. Automático. Computadora. 2020, 17, 562–571. [Referencia cruzada]
16. Tai, Dakota del Norte; Trieu, Nuevo México; Thinh, NT Modelado de posiciones y orientaciones de flores de melón para polinización automática. Agricultura 2024, 14, 746. [Referencia cruzada]
17. Ahmad, K.; Park, J.-E.; Ilyas, T.; Lee, J.-H.; Kim, S.; Kim, H. Polinizaciones precisas y robustas para sandías utilizando inteligencia servovisual guiado. Computadora. Electrón. Agrícola. 2024, 219, 108753. [Referencia cruzada]
18. Yang, M.; Lyu, H.; Zhao, Y.; Sol, Y.; Pan, H.; Sol, Q.; Chen, J.; Qiang, B.; Yang, H. Entrega de polen a pistilos de flores de forsythia. de forma autónoma y precisa mediante un brazo robótico. Computadora. Electrón. Agrícola. 2023, 214, 108274. [Referencia cruzada]
19. Zhou, C. Yolact++ Mejor segmentación de instancias en tiempo real; Universidad de California: Davis, CA, EE. UU., 2020.
20. Carión, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. Detección de objetos de un extremo a otro con transformadores. En Conferencia Europea sobre Visión por Computador; Springer: Cham, Suiza, 2020; págs. 213–229.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gómez, AN; Kaiser, Ł.; Polosukhin, I. Atención es todo lo que necesitas. arXiv 2017, arXiv:1706.03762v7.
22. Jiang, K.; Peng, P.; Lian, Y.; Xu, W. El método de codificación de incrustaciones de posición en un transformador de visión. J. Vis. Comunitario. Representación de imagen . 2022, 89, 103664. [Referencia cruzada]
23. Él, K.; Zhang, X.; Ren, S.; Sun, J. Aprendizaje residual profundo para el reconocimiento de imágenes. En Actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones, Las Vegas, NV, EE. UU., 27 a 30 de junio de 2016; págs. 770–778.
24. Kingma, DP; Ba, J. Adam: Un método para la optimización estocástica. arXiv 2014, arXiv:1412.6980.

electrónica

MDPI

Artículo

Filtros de correlación reprimida de aberración adaptativa con Optimización cooperativa en vehículos no tripulados de alta dimensión Asignación de tareas de vehículos aéreos y planificación de trayectorias

Zijie Zheng [1,*], Zhijun Zhang [1], Zhenzhang Li [2], Qiuda Yu [1] y Ya Jiang [1]

[1] Escuela de Ciencias e Ingeniería de Automatización, Universidad Tecnológica del Sur de China, Guangzhou 510006,
[2] Escuela de Matemáticas y Ciencias de Sistemas de China, Universidad Normal Politécnica de Guangdong, Guangzhou 510665, China; zhenzhangli@gpnu.edu.cn
* Correspondencia: 201910102802@mail.scut.edu.cn

Resumen: En el campo en rápida evolución de las aplicaciones de vehículos aéreos no tripulados (UAV), la complejidad de la planificación de tareas y la optimización de trayectorias, particularmente en entornos operativos de alta dimensión, es cada vez más desafiante. Este estudio aborda estos desafíos mediante el desarrollo del algoritmo de optimización cooperativa del filtro de correlación de supresión de distorsión adaptativa (ARCF-ICO), diseñado para la asignación de tareas y la planificación de trayectorias de vehículos aéreos no tripulados de alta dimensión. El algoritmo ARCF-ICO combina tecnologías avanzadas de filtro de correlación con técnicas de optimización multiobjetivo, mejorando la precisión de la planificación de trayectorias y la eficiencia de la asignación de tareas. Al incorporar las condiciones climáticas y otros factores ambientales, el algoritmo garantiza un rendimiento sólido en altitudes bajas. El algoritmo ARCF-ICO mejora la estabilidad y precisión del seguimiento de UAV al suprimir distorsiones, lo que facilita la selección de ruta óptima y la ejecución de tareas. La validación experimental utilizando los conjuntos de datos UAV123@10fps y OTB-100 demuestra que el algoritmo ARCF-ICO supera a los métodos existentes en área bajo la curva (AUC) y métricas de precisión. Además, la consideración del consumo de batería y la resistencia del algoritmo valida aún más su aplicabilidad a las tecnologías UAV actuales. Esta investigación avanza en la planificación de misiones de vehículos aéreos no tripulados y establece nuevos estándares para el despliegue de vehículos aéreos no tripulados en aplicaciones civiles y militares, donde la adaptabilidad y la precisión son fundamentales.

Palabras clave: optimización multiobjetivo; planificación de trayectorias de vehículos aéreos no tripulados; filtros de correlación; algoritmos adaptativos ; asignación de tareas

Cita: Zheng, Z.; Zhang, Z.; Li, Z.; Yu, Q.; Jiang, Y. Filtros de correlación reprimida de aberración adaptativa con optimización cooperativa en la asignación de tareas y planificación de trayectorias de vehículos aéreos no tripulados de alta dimensión. Electrónica 2024, 13, 3071. https://doi.org/10.3390/

electrónica13153071

Editor académico: Zhiquan Liu

Recibido: 29 de mayo de 2024
Revisado: 20 de julio de 2024
Aceptado: 30 de julio de 2024
Publicado: 2 de agosto de 2024

1. Introducción

En la era del Internet de las cosas, los vehículos aéreos no tripulados (UAV) se han convertido en herramientas esenciales en diversos ámbitos debido a su flexibilidad y eficiencia. A medida que aumentan la complejidad y diversidad de las misiones de vehículos aéreos no tripulados, la necesidad de una asignación de tareas y una planificación de trayectorias avanzadas y multiobjetivos de alta dimensión se vuelve crítica. Estudios recientes de Chen et al. [1–3] han introducido algoritmos innovadores que mejoran la planificación de rutas y el control del comportamiento cooperativo de vehículos aéreos no tripulados heterogéneos. Estos métodos mejoran la eficiencia operativa y la adaptabilidad de los UAV en entornos dinámicos y complejos, abordando los desafíos contemporáneos en el despliegue de UAV.

El rápido desarrollo de la tecnología UAV ha dado lugar a una creciente demanda de aplicaciones UAV en entornos multitarea. Sin embargo, en entornos complejos con múltiples objetivos , un solo UAV a menudo tiene dificultades para manejar múltiples tareas simultáneamente, lo que requiere la colaboración de varios UAV. La investigación sobre la asignación de tareas y la planificación de trayectorias de vehículos aéreos no tripulados multiobjetivos de alta dimensión basada en el aprendizaje profundo tiene como objetivo abordar los desafíos de optimización en la asignación colaborativa de tareas de vehículos aéreos no tripulados múltiples, logrando una gestión y programación inteligentes de los vehículos aéreos no tripulados para mejorar la eficiencia y precisión de la ejecución de tareas

En el ámbito del desarrollo estratégico nacional, la progresión y utilización de la tecnología de vehículos aéreos no tripulados (UAV) emergen como habilitadores tecnológicos fundamentales para iniciativas estratégicas. A través del refinamiento de la planificación de tareas de los UAV, sectores fundamentales dentro de una nación pueden ser testigos de mayores niveles de productividad y eficiencia, fomentando así el cultivo de una productividad novedosa y de alta calidad, un concepto defendido por Yao y Liu [5]. Su trabajo subraya la importancia de delinear marcos teóricos y vías prácticas para impulsar el progreso. Además, la evolución de la tecnología UAV no sólo aumenta las capacidades tecnológicas de una nación sino que también fortalece su competitividad global, como lo ilustran Ma y Chen [6]. Su exploración de las transiciones nacionales subraya el papel indiscutible de la innovación tecnológica en la configuración de las trayectorias económicas y estratégicas. Con la tecnología UAV como piedra angular de la innovación, sus aplicaciones multifacéticas en dominios como la defensa nacional, el transporte y la agricultura contribuyen significativamente a mejorar la competitividad general de una nación. Por lo tanto, al alinear la tecnología de los vehículos aéreos no tripulados con imperativos estratégicos más amplios, las aspiraciones de promover el progreso y fortalecer la competitividad nacional, tal como las propugna el nuevo paradigma de desarrollo, están preparadas para hacerse realidad.

Los modelos de aprendizaje profundo han desempeñado un papel importante en la asignación de tareas y la planificación de trayectorias de los vehículos aéreos no tripulados. Los modelos de aprendizaje profundo comunes incluyen redes neuronales convolucionales (CNN) [7], redes neuronales recurrentes (RNN) [8], aprendizaje por refuerzo profundo (DRL) [9], redes generativas adversarias (GAN) [10] y aprendizaje por transferencia, cada uno de ellos. con sus ventajas y limitaciones. Por ejemplo, las CNN son adecuadas para tareas de procesamiento de imágenes, pero carecen de eficiencia en el manejo de datos secuenciales [11], mientras que las RNN pueden procesar datos secuenciales pero sufren problemas como gradientes que desaparecen y explotan. DRL puede manejar tareas con recompensas retrasadas pero implica un proceso de entrenamiento complejo, mientras que las GAN pueden generar datos pero exhiben inestabilidad durante el entrenamiento. La transferencia de aprendizaje puede aprovechar el conocimiento adquirido previamente para acelerar el aprendizaje en nuevas tareas, pero requiere abordar l

Este estudio tiene como objetivo abordar los problemas de planificación de rutas y asignación de tareas de UAV multiobjetivo de alta dimensión utilizando un filtro de correlación de supresión de distorsión adaptativa (ARCF). Desarrollaremos una red ARCF que integre los estados, acciones y funciones de recompensa de las misiones de vehículos aéreos no tripulados para permitir la toma de decisiones inteligente para la asignación de tareas y la planificación de rutas. Específicamente, emplearemos técnicas de aprendizaje por refuerzo profundo para entrenar la red, permitiéndole aprender estrategias de comportamiento óptimas para vehículos aéreos no tripulados en entornos complejos. Durante el proceso de capacitación, utilizaremos unidades de memoria de repetición para almacenar experiencias históricas y combinaremos redes de objetivos y de estimación para mejorar la estabilidad y eficiencia del proceso de aprendizaje.

Las principales contribuciones de este trabajo son las siguientes:

- Proponer un novedoso algoritmo de optimización cooperativa del filtro de correlación de supresión de distorsión adaptativa (ARCF-ICO), que mejora la precisión y estabilidad de la planificación de misiones de vehículos aéreos no tripulados.
- Integre técnicas de optimización multiobjetivo, logrando una toma de decisiones inteligente y eficiente para la asignación de tareas de UAV y la planificación de rutas en entornos complejos. • Presentar resultados experimentales que muestran que el algoritmo ARCF-ICO supera a los métodos existentes en términos de AUC y métricas de precisión en conjuntos de datos UAV123@10fps y OTB-100.

2. Trabajo relacionado

2.1. Optimización de objetivo único para la planificación de trayectorias de vehículos aéreos no tripulados

La planificación de trayectorias para vehículos aéreos no tripulados (UAV) es fundamentalmente un problema de optimización con implicaciones prácticas. Li y Duan [12] incorporaron el costo de la amenaza y el costo del combustible en un objetivo de optimización ponderado y emplearon un algoritmo de búsqueda gravitacional universal mejorado para mejorar la convergencia de la búsqueda global, mejorando así la calidad de las soluciones óptimas para las trayectorias de los UAV. Qu et al. [13] combinaron un optimizador de lobo gris simplificado con una búsqueda mejorada de organismos simbióticos para proponer un algoritmo híbrido novedoso para obtener rutas factibles y efectivas. Dasdemir et al. [14] diseñó un gene

Algoritmo evolutivo de objetivo único basado en preferencias para optimizar tanto la distancia total de las rutas planificadas como las amenazas de detección de radar. Yao et al. [15] introdujeron un algoritmo híbrido basado en un modelo de control predictivo y un optimizador de lobo gris mejorado para planificar trayectorias óptimas para el seguimiento de objetivos de vehículos aéreos no tripulados en entornos urbanos. Papaioannou et al. [16] abordaron los desafíos de monitorear pasivamente múltiples objetivos en movimiento en entornos obstruidos con vehículos aéreos no tripulados mediante el diseño de un controlador de guía predictiva modelo combinado con una estrategia conjunta de estimación y control. Ren et al. [17] propusieron un enfoque de planificación de rutas multiobjetivo (MOPP) utilizando el algoritmo genético de clasificación no dominado II (NSGA-II), optimizado tanto para la distancia como para la seguridad, demostrando su eficiencia en un entorno urbano mediante el empleo de subdivisiones espaciales basadas en octrees. y mapas de índices de seguridad.

Sin embargo, la investigación actual sobre la planificación de trayectorias para vehículos aéreos no tripulados únicos y múltiples a menudo se centra en la optimización de un solo objetivo, ya sea considerando un solo objetivo o integrando múltiples objetivos de optimización en uno solo mediante ponderación lineal. Dichos enfoques de optimización dependen en gran medida de coeficientes de ponderación subjetivos establecidos por quienes toman las decisiones , lo que impacta directamente en los resultados de la optimización, y pueden pasar por alto trayectorias con un desempeño sobresaliente en objetivos relativamente menores. En los últimos años, a pesar de la creciente atención a los problemas de planificación de trayectorias basados en optimización multiobjetivo, que normalmente consideran sólo dos o tres objetivos optimizados, los requisitos prácticos de optimización para la planificación de trayectorias de vehículos aéreos no tripulados se extienden más allá de un número limitado de objetivos. Para abordar este problema, establecer un modelo de planificación de trayectorias basado en una optimización multiobjetivo de alta dimensión se vuelve particularmente crucial para optimizar simultáneamente varios aspectos de rendimiento de las trayectorias.

2.2. Optimización multiobjetivo de alta dimensión para la planificación de trayectorias de vehículos aéreos no tripulados

Los problemas de optimización multiobjetivo de alta dimensión prevalecen tanto en la vida como en las prácticas de ingeniería, donde múltiples objetivos necesitan optimización simultánea, a menudo con correlaciones interobjetivos que conducen a situaciones conflictivas. En tales casos, es necesario considerar esquemas de optimización alternativos para garantizar la generación de soluciones equivalentes.
en ausencia de información correlacionada de otros esquemas.

Storn y Price [18] propusieron el algoritmo de evolución diferencial (DE), un método basado en la población similar a los mecanismos de reemplazo en estado estacionario, para resolver problemas de optimización de parámetros reales. Los nuevos descendientes solo compiten con sus padres correspondientes, y si los descendientes presentan una mejor aptitud física, los reemplazan. Con el surgimiento de nuevas heurísticas bioinspiradas, como la optimización del enjambre de partículas [19], el algoritmo del lobo gris [20], el algoritmo de la ballena [21] y el algoritmo de búsqueda del gorrión [22], es posible comprender cómo se aplican a diferentes tipos de problemas de optimización objetiva. se vuelve crucial. La literatura ha optimizado la distancia de la trayectoria del UAV y el costo de la amenaza de la trayectoria utilizando algoritmos genéticos y los ha suavizado [23].

La demarcación de los algoritmos de optimización multiobjetivo de alta dimensión radica en si el número de objetivos optimizados supera los cuatro [24]. Con un número cada vez mayor de objetivos optimizados, el número de soluciones no dominadas generadas durante el proceso de resolución de algoritmos de optimización multiobjetivo aumenta exponencialmente, afectando significativamente el rendimiento y la eficiencia del algoritmo [25,26]. Además, la presión de selección generada por los algoritmos de optimización multiobjetivo al resolver problemas multiobjetivo de alta dimensión suele ser insuficiente para guiar a los individuos de la población hacia puntos ideales.

Las estrategias para mejorar estos dos indicadores se dividen principalmente en tres categorías: (1) Mejorar la presión de selección de los algoritmos cambiando los métodos de dominancia de Pareto para acelerar la tasa de convergencia de las poblaciones. GrEA [27] utiliza métricas de evaluación basadas en cuadrículas para mejorar la presión de selección de los algoritmos. NSGA-III [28] utiliza una estrategia de punto de referencia en lugar de la estrategia de distancia de aglomeración de NSGA-II [29] para seleccionar individuos excelentes de soluciones no dominadas, mejorando así la convergencia de los algoritmos. 1by1EA [30] selecciona descendientes individuales uno por uno basándose en la convergencia individual cuando ocurre la selección ambiental, luego mejora la diversidad de poblaciones a través de técnicas de nicho. RPEA [31] continuamente

Genera una serie de puntos de referencia bien convergentes y distribuidos basados en la población actual para guiar la evolución. (2) Descomponer un problema complejo de optimización multiobjetivo de alta dimensión en un grupo de subproblemas y cooptimizar estos subproblemas.
MOEA [32] descompone el problema en una serie de subproblemas de optimización de un solo objetivo descompuestos adaptativamente y luego agrega información de problemas vecinos. MaOEA/Ds [33] utiliza un conjunto de vectores de referencia autoguiados distribuidos uniformemente en el espacio para dividir el espacio de decisión en múltiples subespacios pequeños y juzgar los méritos de los individuos en los subespacios. Yi et al. [34] propusieron un método de optimización evolutiva multiobjetivo basado en la descomposición objetiva, descomponiendo el problema en varios subproblemas y resolviendo cada subproblema en paralelo, utilizando completamente la información de otras subpoblaciones para mejorar la presión de selección de Soluciones no dominadas. (3) Estrategia basada en métricas de evaluación. Evaluar la superioridad e inferioridad de los individuos de manera integral a través de métricas de evaluación y luego seleccionar individuos excelentes para operaciones genéticas. Sin embargo, la complejidad computacional de tales estrategias suele ser alta y dichas métricas de evaluación se utilizan comúnmente para evaluar la calidad de los resultados de optimización de algoritmos [35].

En el proceso de planificación colaborativa de trayectorias para múltiples UAV, es necesario considerar no solo los atributos de trayectoria individuales sino también la coordinación espacial entre múltiples UAV. Para evitar el impacto de combinar múltiples objetivos de optimización en uno solo mediante la ponderación en la planificación de trayectorias, se propone un modelo basado en optimización multiobjetivo de alta dimensión para la planificación colaborativa de trayectorias para múltiples UAV. Este modelo optimiza el costo de la distancia de la trayectoria del UAV, el costo de seguridad de la trayectoria, el costo de la energía de la trayectoria y la coordinación espacial entre múltiples UAV como objetivos de optimización. A diferencia de los enfoques existentes que tratan la trayectoria de un solo UAV como un individuo en la población y optimizan las trayectorias de múltiples UAV por separado, en este modelo, las trayectorias de múltiples UAV se tratan como una entidad completa en la población y la optimización se realiza simultáneamente. en múltiples trayectorias de UAV. Además, se llevan a cabo evaluaciones integrales de la convergencia y diversidad de individuos en la población, y se mejoran las estrategias de apareamiento del algoritmo para problemas de planificación de trayectorias para mejorar el rendimiento de la convergencia [35,36]. A través del algoritmo, se obtiene un conjunto de trayectorias óptimas de Pareto para múltiples UAV para que las utilicen los tomadores de decisiones. Los tomadores de decisiones pueden seleccionar las trayectorias más adecuadas para los atributos de su misión de este conjunto de trayectorias

3. Modelado de planificación de rutas de vuelo de múltiples UAV 3.1.

Descripción del problema En el contexto de la planificación colaborativa de trayectorias para múltiples vehículos aéreos no tripulados (UAV), un grupo de UAV tiene la tarea de navegar desde múltiples puntos de partida hasta una serie de puntos objetivo específicos para ejecutar misiones complejas. El escenario de la misión se establece dentro de un área protegida por varios sistemas de defensa, donde los UAV deben evadir inteligentemente las amenazas en la cobertura de radar enemigo y las zonas de fuego antiaéreo, considerando al mismo tiempo las limitaciones de rendimiento y los requisitos de cooperación de cada UAV. Con base en este escenario, llevamos a cabo análisis de simulación de rutas de ejecución de misiones y navegación de grupos de UAV en un espacio tridimensional. Los supuestos computacionales son los siguientes:

(1) Todos los UAV mantienen una velocidad de vuelo constante durante la ejecución de la misión. (2) Cada segmento de trayectoria se divide en trayectorias de vuelo rectas. (3) Cada UAV posee características de rendimiento idénticas.

El núcleo de la planificación colaborativa de trayectorias para múltiples UAV implica el diseño de modelos de costos de trayectoria y modelos espaciales colaborativos. El costo de la trayectoria considera principalmente la distancia, las amenazas y el consumo de energía. El objetivo del modelo es minimizar estos costos y al mismo tiempo mejorar la coordinación espacial entre los UAV.

Suponiendo que hay Nz puntos objetivo, cada uno de los cuales requiere tareas de reconocimiento, ataque y confirmación representados como Smj , donde j [1, Nz] denota el tipo de tarea m del j-ésimo punto objetivo , correspondiente a reconocimiento, ataque y confirmación (m = 1, 2, 3). Por lo tanto, el número total de tareas es $\sum_{norte=1}^{Número\ Objetivo} 3m = 1Smn$. Cada tarea tiene ventanas de tiempo específicas [cmn , dmn], $\sum$ los requisitos de munición designados zj para las tareas de ataque en cada punto objetivo.

El grupo de vehículos aéreos no tripulados está formado por vehículos aéreos no tripulados de reconocimiento Np , vehículos aéreos no tripulados de ataque Nq y vehículos aéreos no tripulados integrados de reconocimiento y ataque Npq , por un total de Nv = Np + Nq + Npq , indexados como u [1, Nv]. Cada UAV u tiene un vector de capacidad Capabilityu correspondiente a su tipo de tarea, donde Capabilityu(m) representa la capacidad del UAV u para realizar tareas de tipo m, tomando valores de 1 o 0. La capacidad de carga útil de cada UAV u es zu, y si El UAV u carece de capacidad de ataque, entonces zu = 0. Se supone que la relación entre el consumo de combustible f pu y la velocidad de vuelo Vu de cada UAV u es f pu = αu × Vu, donde α representa la proporcionalidad entre el consumo de combustible y la velocidad de vuelo. , y Vu está dentro del rango [ku, ju], que denota las velocidades mínima y máxima.

En entornos urbanos, el problema de múltiples UAV que rastrean múltiples objetivos terrestres se muestra en la Figura 1, donde N UAV están listos para ejecutar M tareas. El conjunto de UAV se denota como U = {U1, U2 ,. . . , UN}, y el conjunto de objetivos se denota como H = {H1, H2, . . . , HM}, con tipos de tareas representados por Mi = {1, 2} (Mi = 1 para reconocimiento, Mi = 2 para ataque).

Para demostrar la heterogeneidad de los UAV y los requisitos de rendimiento específicos de los escenarios de misión, se utiliza Zum para representar la matriz de rendimiento de los UAV, donde los elementos de la matriz representan la capacidad del UAV para ejecutar una determinada tarea. Por ejemplo, si hay tres UAV y su matriz de rendimiento Zum se muestra en la Tabla 1, Zum (1, 1) = 0,9 indica que la capacidad de la tarea de reconocimiento del UAV 1 es 0,9. Si H1 es una tarea de reconocimiento (M1 = 1) con una capacidad mínima requerida de 0,6, entonces entre los tres UAV U1, U2, U3, solo U1 (0,9) puede cumplir los requisitos de la tarea de H1.

Tabla 1. Valor de capacidad del UAV.

Vehículo aéreo no tripulado	Valor de habilidad	
	Explorar	Pista
U1	0,9	0,4
U2	0.3	0,9
U3	0,5	0,5

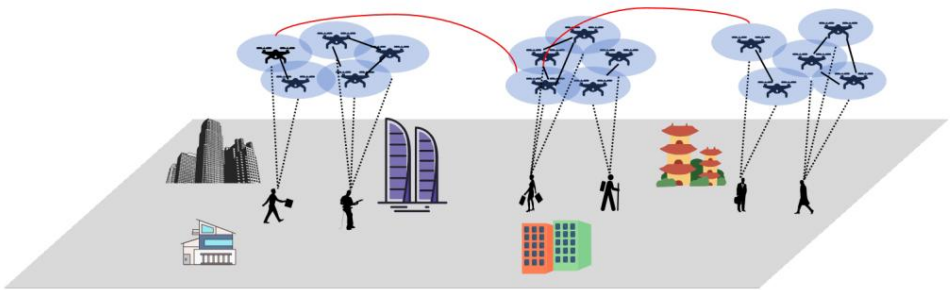


Figura 1. Asignación de tareas de múltiples UAV en escenarios urbanos.

3.2. Modelo objetivo de diseño de optimización

Nuestro objetivo es minimizar el consumo total de combustible de toda la flota de UAV durante la ejecución de la tarea y al mismo tiempo maximizar la eficiencia de la finalización de la tarea. Por lo tanto, definimos las siguientes funciones de costos.

Costo Total del Consumo de Combustible: Representa el consumo total de combustible de los UAV desde el despegue, ejecución de la tarea, hasta el regreso al aeropuerto. El consumo de combustible de cada UAV depende de la distancia y la velocidad del vuelo. Si definimos el consumo de combustible del UAV u que realiza la tarea k desde el punto Pi al Pj como fu,k,i,j , entonces el consumo total de combustible Ff uel se puede expresar como

$$_{\text{Combustible}} = \sum_{u=1}^{Nu} \sum_{k=1}^{3} \sum_{i=0}^{\text{Número de puntos}} \sum_{j=1}^{\text{Número de puntos}} X_{u,k,i,j} \cdot f_{u,k,i,j} \quad (1)$$

Costo de aptitud del tiempo de tarea: Mide la diferencia entre el tiempo de finalización de las tareas y el punto medio de sus ventanas de tiempo. El objetivo es hacer que la tarea de los UAV

tiempos de ejecución lo más cerca posible del punto medio de las ventanas de tiempo de la tarea. Si denotamos la aptitud de la ventana de tiempo del UAV u como Tadapt,u, entonces la aptitud de la ventana de tiempo total Ftime es

$$\text{Ftiempo} = \sum_{tu=1}^{Nu} \text{Tadapt,u} = \sum_{tu=1}^{Nu} \sum_{yo=0}^{\_\_} \sum_{j=1}^{\_\_} \sum_{k=1}^{3} \text{Xu,k,i,j} \cdot (\text{tu,k,i,j} - \frac{\text{sk,j} + \text{ek,j}}{2}) \quad (2)$$

Combinando las dos funciones de costos anteriores, formamos un problema de optimización bioobjetivo:

$$\text{Minimizar F} = \alpha \cdot \text{Ff uel} + \beta \cdot \text{Ftime} \quad (3)$$

donde α y β son parámetros que equilibran la importancia de los dos objetivos. En la Ecuación (3), las funciones objetivo se normalizan para garantizar que cada una contribuya por igual a la optimización general. El proceso de normalización implica escalar cada función objetivo a un rango [0, 1] en función de sus respectivos valores máximo y mínimo observados durante las ejecuciones iniciales.

3.3. Restricciones

En el problema de planificación de trayectorias de vehículos aéreos no tripulados, debemos considerar las siguientes limitaciones, incluidas las condiciones climáticas para tener en cuenta las operaciones a baja altitud.

Restricción de ejecución de tareas: asegúrese de que cada tarea sea ejecutada por al menos un UAV con la capacidad correspondiente. La fórmula de restricción para la ejecución de tareas es la siguiente:

$$\sum_{tu=1}^{Nu} \text{Xu,k,i,j} \geq 1 \quad \text{k,} \quad \text{i,} \quad \text{j} \quad (4)$$

Aquí, Xu,k,i,j representa la variable de decisión binaria que indica si el UAV u ejecuta la tarea k desde el punto Pi hasta Pj . Nu es el número total de vehículos aéreos no tripulados. k representa el tipo de tarea. i representa el índice del punto de partida. j representa el índice del punto final.

Restricción de coincidencia de capacidad: las tareas realizadas por los UAV deben cumplir con sus limitaciones de capacidad. La fórmula de restricción para la coincidencia de capacidades es la siguiente:

$$\sum_{tu=1}^{Nu} \text{Xu,k,i,j} \geq 1 \quad \text{k,} \quad \text{i,} \quad \text{jXu,k,i,j} \leq \text{Capacidadu,k} \quad \text{u,} \quad \text{k,} \quad \text{i,} \quad \text{j} \quad (5)$$

donde Capacidadu,k representa la capacidad del UAV u para realizar la tarea k.

Restricción de las condiciones climáticas: las operaciones a bajas altitudes están influenciadas por las condiciones climáticas como la velocidad del viento, las precipitaciones y la visibilidad. Estos factores se incorporan en la estimación de la trayectoria para garantizar operaciones seguras y confiables. La fórmula de restricción para las condiciones climáticas es la siguiente:

$$\text{Wu,k,i,j} \leq \text{Wmáx} \quad \text{u,} \quad \text{k,} \quad \text{i,} \quad \text{j} \quad (6)$$

donde Wu,k,i,j representa el impacto climático en el UAV u mientras realiza la tarea k desde el punto Pi hasta Pj . Wmax es el impacto climático máximo permitido.

Restricción de tiempo de vuelo: asegúrese de que el tiempo para que un UAV vuele de un punto de tarea al siguiente no exceda el valor máximo especificado. La fórmula de restricción para el tiempo de vuelo es la siguiente:

$$\text{tu,k,i,j} - \text{tu,k,i,j}-1 \leq \text{Tmax} \quad \text{u,} \quad \text{k,} \quad \text{i,} \quad \text{j} > 1 \text{ donde} \quad (7)$$

tu,k,i,j representa el momento en el que el UAV u llega a la tarea punto j mientras se realiza la tarea k, comenzando desde el punto Pi . Tmax es el tiempo de vuelo máximo permitido entre puntos de tarea consecutivos.

3.4. Métricas de desempeño

La evaluación del desempeño es crucial para validar la efectividad de nuestro modelo de optimización. Utilizamos las siguientes métricas de rendimiento:

Cobertura de tareas: la proporción de tareas asignadas y completadas con éxito con respecto al número total de tareas. La cobertura de tareas se puede expresar como

$$\text{Cobertura} = \frac{\text{Número de tareas asignadas y completadas con éxito}}{\text{Número total de tareas}} \quad (8)$$

Consumo promedio de batería: el consumo promedio de batería de la formación de UAV para realizar todas las tareas se expresa de la siguiente manera:

$$\text{Consumo promedio de batería} = \frac{\sum_{u=1}^{Nu} \text{Batería usada por ti}}{Nu} \quad (9)$$

Resistencia: La resistencia mide el tiempo operativo de los UAV, asegurando que puedan completar tareas dentro de los límites de la batería. Se expresa como

$$\text{Resistencia} = \frac{\sum_{u=1}^{Nu} \text{Tiempo operativo de u}}{Nu} \quad (10)$$

Eficiencia del tiempo: la diferencia entre el tiempo de finalización de todas las tareas y el hora de inicio más temprana. Se puede expresar como

$$\text{Eficiencia de tiempo} = \max_{u \in U} \{t\text{completado},u\} - \min_{u \in U} \{t\text{start},u\} \quad (11)$$

A través de estas métricas de rendimiento, podemos evaluar de manera integral la efectividad del algoritmo de optimización, permitiendo ajustes adicionales a los parámetros del modelo o mejoras algorítmicas.

4. Algoritmo de estimación multiobjetivo genético adaptativo El algoritmo ARCF

(Filtro de corrección del mapa de respuesta adaptativa) tiene como objetivo integrar la distorsión del mapa de respuesta que ocurre durante el proceso de seguimiento con el proceso de entrenamiento del filtro, mejorando así el rendimiento del algoritmo (como se muestra en la Figura 2).). Para suprimir la distorsión del mapa de respuesta, el primer paso es la identificación de la distorsión (es decir, determinar cuándo se produce la distorsión del mapa de respuesta). Introduce la norma euclidiana para definir la diferencia entre los mapas de respuesta del cuadro anterior M1 y el cuadro actual M2 .

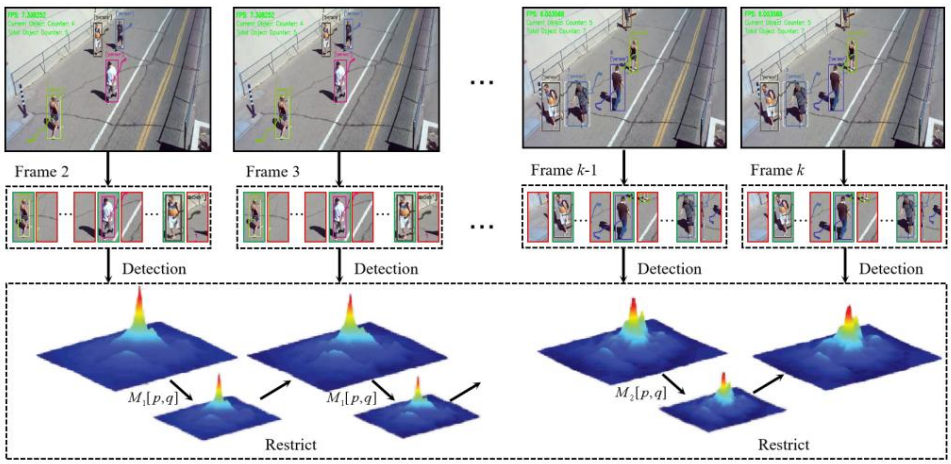


Figura 2. Diagrama de flujo del algoritmo ARCF.

El objetivo principal del algoritmo ARCF es combinar la distorsión del mapa de respuesta generada durante el seguimiento del objetivo con el proceso de entrenamiento del filtro, mejorando el rendimiento del seguimiento mediante la actualización dinámica del filtro. La distorsión del mapa de respuesta es causada principalmente por el movimiento rápido del objetivo o factores ambientales externos como cambios de oclusión y iluminación. El algoritmo ARCF identifica y suprime estas distorsiones analizando los cambios en los mapas de respuesta entre fotogramas consecutivos, específicamente calculando la distancia euclidiana entre los mapas de respuesta de dos fotogramas.

En el algoritmo, los mapas de respuesta de dos cuadros consecutivos F1 y F2 se denominan M1 y M2, respectivamente, y se alinean utilizando la operación de desplazamiento ψ para calcular sus diferencias. La fórmula para calcular la diferencia es

$$\Delta M = \psi(M1, M2) = \sum_{pqq} \overline{(M1[p, q] - M2[p, q])2} \quad (12)$$

Aquí, p y q representan las coordenadas espaciales del mapa de respuesta.

Con base en la medida de diferencia antes mencionada, la función objetivo del algoritmo ARCF puede describirse como un problema de optimización, cuyo objetivo es minimizar la distorsión del mapa de respuesta mientras se maximiza la precisión del seguimiento del objetivo. La función objetivo consta de un término de distorsión y un término de regularización, expresado como

$$\min h \quad \lambda \quad h \quad ^2 + \gamma \quad \sum_{re=1}^{D} h \quad Md - Yd \quad ^2 + \quad \Delta M \quad (13)$$

donde h es el filtro, denota la operación de convolución, Md es el mapa de respuesta de entrada para el d-ésimo canal, Yd es el mapa de respuesta de salida deseada y λ, γ y son coeficientes que ajustan la importancia de cada término.

Para lograr eficiencia computacional, la función objetivo se transforma aún más en la dominio de la frecuencia. En el dominio de la frecuencia, la fórmula se convierte en

$$\min_{h} \quad \lambda \quad H \quad ^2 + \gamma \quad F(H) \quad X - Y \quad ^2 + \quad \Delta M \quad (14)$$

donde F representa la transformada de Fourier, denota multiplicación por elementos, y X e Y son las representaciones del dominio de frecuencia de entrada y salida deseada, respectivamente.

Al emplear el algoritmo ADMM (Método de multiplicadores de dirección alterna) para resolver el problema de optimización, el filtro se puede actualizar de manera efectiva. El proceso de solución implica dos subproblemas principales: optimizar el filtro y actualizar el mapa de respuesta. Este enfoque suprime eficazmente la distorsión del mapa de respuesta causada por movimientos rápidos o cambios ambientales externos, mejorando así la estabilidad y precisión del algoritmo de seguimiento. El algoritmo ARCF se puede resumir como Algoritmo 1.

Algoritmo 1: El procedimiento de ARCF

Inicializar filtro h0, tasas de aprendizaje λ, γ,
Establecer iteraciones máximas
T para t = 1 a T hacer
  Capturar fotograma actual Ft
  Calcule el mapa de respuesta Mt usando ht−1
  si t > 1 entonces
    Calcular la distorsión $\Delta M = \sum_{p,q} \overline{(Mt[p, q] - Mt-1[p, q])^2}$
  terminara si
  Actualizar filtro en el dominio de frecuencia:
  Ht = arg minH λ H $^2 + \gamma$ F(H) Xt − Yt $^2 +$ ΔM
  Actualizar mapa de respuesta Mt :
  Mt = Ht Pies
  Verifique la convergencia:
  si F(Ht) − F(Ht−1) < umbral entonces
    romper
  el final si
  Actualice ht con Ht en el extremo del dominio
espacial para
Genere la mejor solución global Gbest = hT

4.1. Algoritmo ARCF-ICO El

algoritmo ARCF-ICO integra mejoras significativas a la estrategia ARCF original, enfocándose tanto en la convergencia como en la diversidad de soluciones dentro de un espacio de optimización de alta dimensión . Esto es particularmente pertinente en el contexto de la planificación de misiones de vehículos aéreos no tripulados , donde los diversos entornos operativos y los requisitos de misión en rápida evolución requieren un enfoque de optimización sólido y adaptable.

En la fase inicial del algoritmo ARCF-ICO, se da prioridad a lograr altas tasas de convergencia. Esto garantiza una rápida alineación de los UAV hacia trayectorias o conjuntos de soluciones óptimos, abordando de manera efectiva las necesidades operativas inmediatas, como la vigilancia o la detección de amenazas. A medida que avanza el algoritmo, el énfasis se desplaza hacia la preservación de la diversidad entre las soluciones. Esto es crucial en las operaciones de vehículos aéreos no tripulados para explorar una variedad de posibles rutas de vuelo o estrategias tácticas, evitando así los óptimos locales y mejorando la solidez de los resultados de la misión. El algoritmo ARCF-ICO está diseñado para ser implementado por varias categorías de UAV, incluidos UAV de ala fija, de ala giratoria e híbridos. Estos UAV se pueden utilizar tanto en aplicaciones comerciales como militares, según sus capacidades y requisitos de la misión.

El indicador de evaluación integral para la convergencia y la diversidad (CAD) [37] es una métrica recientemente propuesta diseñada para evaluar tanto la convergencia como la diversidad de soluciones dentro de nuestro marco ARCF-ICO. Este novedoso indicador se define de la siguiente manera:

$$CAD(ui , U) = 1 + rand(0,8, 1) \times M \times \left(\frac{t}{tmáx}\right)^{\theta} \times D(ui , U) \times (1 - C(ui , U)) \quad (15)$$

donde D(ui , U) representa la medida de diversidad y C(ui , U) denota la medida de convergencia para UAV ui dentro de la flota U. El parámetro M denota el número de objetivos, y θ rige el equilibrio entre convergencia y diversidad como el El algoritmo itera desde t hasta tmax, el número máximo de generaciones.

La diversidad D(ui , U) se calcula como

$$SDE(ui , U) = \min_{uj \in U, j \neq i} \sqrt{\sum_{k=1}^{metro} sde(f_k'(ui , U), f_k(uj', U))^2} \quad (dieciséis)$$

donde $sde(f_k'(ui , U), f_k(uj', U))$ se define como:

$$sde(f)_k'(ui , U), f_k(uj', U)) = \begin{cases} f_k'(uj , U) - f_k'(ui , U) & f_k(uj', U) > f \text{ en caso contrario } {}_k'(ui , U) \\ 0 & \end{cases} \quad (17)$$

La convergencia C(ui , U) de un UAV en relación con la flota se cuantifica como

$$C(ui , U) = \sqrt{\frac{Disco\ (ui , U)}{metro}} \quad (18)$$

donde Disc(ui , U) representa la distancia euclidiana desde el UAV ui hasta la solución ideal en el espacio objetivo normalizado.

Al integrar estas estrategias, ARCF-ICO adapta dinámicamente la planificación de misiones de vehículos aéreos no tripulados y las tácticas de respuesta de acuerdo con las condiciones ambientales en evolución y las demandas operativas, optimizando tanto la eficiencia como la eficacia de los vehículos aéreos no tripulados desplegados.

4.2. Planificación multiobjetivo ARCF-ICO

En el modelo de optimización multiobjetivo ARCF-ICO, hemos ideado una estrategia de acoplamiento eficiente adaptada a los entornos de tareas complejas que encuentran los UAV. Esta estrategia combina los rasgos favorables de los individuos dentro de la población e introduce elementos estocásticos para aumentar la diversidad de la población, mejorando así la adaptabilidad y flexibilidad del algoritmo. El algoritmo ARCF-ICO es aplicable a misiones realizadas en Visual Line Of Sight (VLOS) [38], Beyond Visual Line Of Sight (BVLOS) [38] y completame

operaciones autónomas. La flexibilidad del algoritmo le permite adaptarse a diferentes restricciones y requisitos operativos.

En el contexto de la planificación de la trayectoria del UAV, los puntos de trayectoria generados por cada Los UAV para cada segmento de trayectoria están representados por la siguiente matriz:

$$X = [X1, X2, \ldots, Xm]^{t} \quad (19)$$

donde Xi es una matriz N × 3L que representa el iésimo individuo de la población, N es el número de vehículos aéreos no tripulados y L es el número de segmentos de trayectoria.

Durante el proceso de apareamiento en la población de padres, se seleccionan dos padres, P1 y P2, en función de su indicador de evaluación integral CAD. La fórmula de cálculo CAD es la siguiente:

$$CAD(P) = \frac{1}{norte}\sum_{yo=1}^{norte} 1 + rand(0,8, 1) \times M \times \left(\frac{t}{tmáx}\right)^{\theta} \times D(Pi , P) \times (1 - C(Pi , P)) \quad (20)$$

donde D(Pi , P) y C(Pi , P) representan los indicadores de diversidad y convergencia de Pi individual , M es el número de funciones objetivo, θ es el parámetro de equilibrio, t es la generación actual y tmax es la máxima generación.

La operación de apareamiento es la siguiente: combina la información de la trayectoria de los dos padres para generar nuevas trayectorias de descendencia:

$$Pnew,j = \alpha P1,j + (1 - \alpha)P2,j \text{ donde} \quad (21)$$

P1,j y P2,j denotan las coordenadas de posición de los padres P1 y P2 en el jésimo punto de la trayectoria, respectivamente, y α es un número aleatorio entre 0 y 1 se utiliza para controlar la proporción de contribución de los dos padres en la descendencia recién generada. La operación de acoplamiento individual se muestra en la Figura 3.

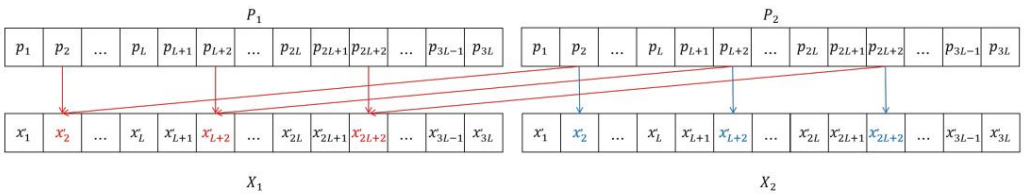


Figura 3. Operación de apareamiento individual.

Este diseño garantiza que los UAV puedan ajustar sus estrategias de vuelo de acuerdo con los requisitos de las tareas reales y los cambios ambientales al ejecutar tareas como vigilancia, reconocimiento u otras misiones complejas. Además, el enfoque de optimización multiobjetivo permite a los UAV optimizar otras métricas de tareas importantes, como el tiempo de vuelo y la eficiencia del combustible, al tiempo que garantiza la eficiencia de las tareas, garantizando así una ejecución integral y de alto rendimiento de la misión.

5. Simulación y Análisis de Resultados 5.1. Conjuntos de datos

Para evaluar exhaustivamente el rendimiento del algoritmo, utilizamos dos conjuntos de datos ampliamente utilizados: el conjunto de datos UAV123@10fps [39] y el conjunto de datos OTB-100 [40]. A continuación se muestran los detalles específicos de cada conjunto de datos:

Conjunto de datos UAV123@10fps: este conjunto de datos comprende 123 escenarios de seguimiento capturados utilizando vehículos aéreos no tripulados en entornos aéreos. Incluye una combinación de escenas del mundo real y sintéticas generadas mediante simuladores. El conjunto de datos cubre 12 entornos de desafíos de seguimiento diferentes, lo que proporciona un conjunto diverso de escenarios para la evaluación.

Conjunto de datos OTB-100: El conjunto de datos OTB-100 consta de 100 escenarios de seguimiento del mundo real capturados manualmente. Abarca 11 entornos distintos de desafíos de seguimiento y ofrece una amplia gama de escenarios para evaluar el rendimiento del algoritmo.

La utilización de estos conjuntos de datos permite una evaluación integral del algoritmo. a través de diversos desafíos de seguimiento y condiciones ambientales. Su información específica se muestra en la Tabla 2 a continuación.

Tabla 2. Detalles del conjunto de datos UAV123@10fps y OTB-100.

UAV123@10fps			OTB-100	
Número de serie	Nombre del desafío	Número	Nombre del desafío	Número
1	Variación de escala (SV)	109	Variación de escala (SV)	[illegible]
2	Cambio de relación de aspecto (ARC)	68	Oclusión (OCC)	49
3	Desorden de fondo (BC)	21	Variación de iluminación (IV)	38
4	Movimiento de la cámara (CM)	70	Desenfoque de movimiento (MB)	31
5	Movimiento rápido (FM)	28	Deformación (DEF)	43
6	Oclusión completa (FOC)	33	Movimiento rápido (FM)	43
7	Variación de iluminación (IV)	31	Rotación fuera del plano (OPR)	64
8	Baja resolución (LR)	48	Rotación en el plano (IPR)	52
9	Fuera de la vista (OV)	30	Desordenes de fondo (BC)	33
10	Oclusión Parcial (POC)	73	Fuera de la vista (OV)	14
11	Objeto similar (SOB)	39	Baja resolución (LR)	10
12	Cambio de punto de vista (VC)	60	-	-

5.2. Configuración de parámetros experimentales

En la configuración de los experimentos, la configuración de hardware utilizada incluye un procesador Intel Core i9-13700 y 32 GB de memoria. La configuración del software es Basado en la plataforma MATLAB R2019a. En cuanto a la configuración de parámetros, la regularización El parámetro se establece en 1.2, siguiendo la configuración del algoritmo ARCF original. Después Para realizar ajustes experimentales, otro parámetro de regularización se establece en 0,001. El aprendizaje La tasa para la plantilla objetivo, denotada por η, se establece en 0,0192, siguiendo nuevamente las pautas del algoritmo ARCF original.

5.3. Índices de evaluación de correlación

Para el algoritmo de optimización multiobjetivo ARCF-ICO en la planificación de misiones UAV, Empleamos tres métricas de evaluación mejoradas para evaluar de manera integral el rendimiento del algoritmo : Área bajo la curva (AUC), Error de ubicación central (CLE) y Precisión. Estos Las métricas miden eficazmente el rendimiento y la precisión de los UAV durante la ejecución de la misión.

Área bajo la curva (AUC): esta métrica evalúa la tasa de éxito general de los UAV en múltiples tareas de vuelo, particularmente en el mantenimiento de objetivos en entornos complejos . El AUC se calcula como

$$\text{ABC} = \frac{\text{área}(\text{Bpred} \cap \text{Bverdadero})}{\text{área}(\text{Bpred} \quad \text{Bverdadero})}, \tag{22}$$

donde Bpred representa el rectángulo de ubicación del objetivo previsto por el algoritmo UAV, y Btrue representa el rectángulo de la verdadera ubicación del objetivo.

Error de ubicación central (CLE): esta métrica mide la distancia euclidiana promedio entre la posición central prevista del UAV y la verdadera posición central del objetivo. La alta precisión en CLE es esencial para garantizar la ejecución precisa de tareas como la supervisión. y reconocimiento. CLE se calcula como

$$\text{CLE} = \sqrt{(\text{xpred} - \text{xverdadero})^2 + (\text{ypred} - \text{yverdadero})^2}, \tag{23}$$

donde (xpred, ypred) es la posición central prevista por el UAV y (xtrue, ytrue) es la posición real posición central del objetivo.

Precisión: La precisión mide la precisión de la localización del UAV dentro de un umbral específico t, es decir, la proporción de cuadros donde el error de predicción de la posición central del objetivo está dentro del umbral. Esta es una métrica fundamental para evaluar el rendimiento del seguimiento en tiempo real de los UAV. La precisión se calcula como

$$\text{Precisión} = \frac{Nt \leq umbral}{total}, \tag{24}$$

donde Nt≤threshold es el número de fotogramas donde CLE es menor o igual que el umbral t, y Ntotal es el número total de fotogramas. El umbral t normalmente se establece en 20 píxeles.

A través de estas tres métricas de evaluación, se puede evaluar de manera integral el desempeño de los UAV en la ejecución de tareas en entornos complejos, como la precisión y la estabilidad de la planificación de rutas. Estas métricas no sólo ayudan a optimizar las estrategias operativas de los UAV, sino que también proporcionan información crucial para mejorar aún más los algoritmos y ajustar los parámetros de vuelo.

5.4. Estudio comparativo

En este capítulo, el algoritmo ARCF-ICO se evalúa utilizando dos conjuntos de datos: UAV123@10fps y OTB-100. Para comprender mejor el rendimiento del algoritmo ARCF-ICO propuesto, se comparará con ocho algoritmos populares en el campo del seguimiento de objetos de vídeo, incluidos KCF [41], LDES [42], MCCT-H [43], Staple [44], fDSST [45] y AutoTrack [46]. Los algoritmos LDES y AutoTrack se publicaron en AAAI2019 y CVPR2020, respectivamente, y se centran en conjuntos de datos de UAV, mientras que los cinco algoritmos restantes son algoritmos de seguimiento clásicos basados en filtros de correlación de los últimos años.

La Figura 4 ilustra la comparación completa del algoritmo ARCF-ICO con los otros seis algoritmos principales en el conjunto de datos UAV123@10fps en términos de AUC y precisión. En la figura, se puede observar que el algoritmo ARCF-ICO propuesto logra el mayor rendimiento, con un AUC integral de 0,516 y una precisión integral de 0,712, superando a los otros seis algoritmos. Los algoritmos AutoTrack y LDES ocupan el segundo y tercer lugar, con valores completos de AUC de 0,504 y 0,492, y valores completos de precisión de 0,682 y 0,655, respectivamente. En comparación con el algoritmo ARCF- ICO, los algoritmos AutoTrack y ARCF tienen un AUC más bajo en un 1,2% y un 2,4%, y una precisión más baja en un 3,0% y un 5,7%, respectivamente. Los cinco algoritmos restantes, MCCT-H, Staple, fDSST y KCF, tienen valores de AUC más bajos, que van de 0,285 a 0,456, y valores de precisión más bajos, que van de 0,384 a 0,581, en comparación con el algoritmo ARCF-ICO.

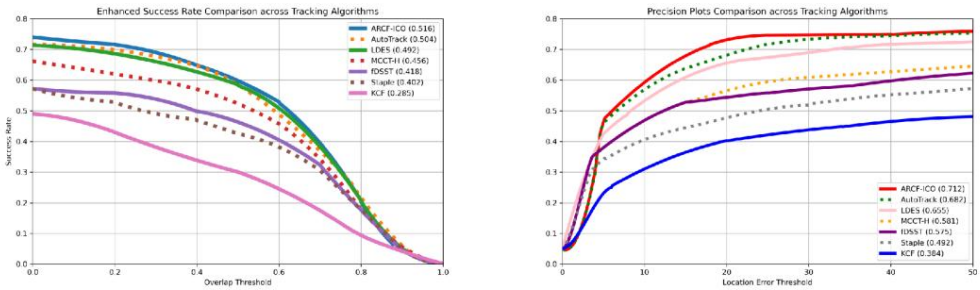


Figura 4. Comparación entre AUC y Precisión en el conjunto de datos UAV123@10fps.

La Tabla 3 presenta la comparación completa de precisión de varios algoritmos en los 12 entornos desafiantes del conjunto de datos UAV123@10fps. Se puede observar en la Tabla 3 que el algoritmo ARCF-ICO ocupa consistentemente el primer lugar en precisión en los 12 entornos desafiantes. Los algoritmos AutoTrack y LDES ocupan el segundo lugar en diez y dos entornos desafiantes, respectivamente. Por lo tanto, los resultados de la comparación de la Tabla 3 demuestran que el algoritmo ARCF-ICO se adapta bien a los desafíos de seguimiento más complejos y exhibe un rendimiento de seguimiento sólido.

Tabla 3. Tabla de comparación de precisión de UAV123@10fps para varios escenarios de desafío del conjunto de datos.

Nombre del desafío	Nuestro algoritmo	Algoritmo de contraste					
	ARCF-ICO	AutoTrack	LDES	MCCT-H	fDSST	Grapa	KCF
SV	0,724	0,672	0,642	0.545	0.521	0,496	0.372
ARCO	0,692	0,686	0.631	0.591	0.552	0.503	0.336
[illegible]	0.702	0.621	0.603	0.503	0.521	0.467	0.415
CM	0,684	0,691	0.633	0.512	0,488	0,472	0.375
FM	0,696	0.664	0,628	0,498	0.508	0,475	0.311
FOC	0,693	0,677	0,689	0.582	0.571	0.436	0.388
IV	0,669	0,673	0,642	0.603	0,598	0,479	0.406
LR	0,688	0.645	0,667	0,654	0.635	0.527	0.416
VO	0,741	0.714	0,656	0,672	0.626	0,426	0.403
POS	0.744	0,725	0.702	0,625	0.645	0.539	0.369
SOLLOZO	0,752	0.714	0,693	0.564	0.586	0.545	0.388
VC	0.763	0.702	0,671	0,625	0.645	0.535	0,426

La Tabla 4 muestra los resultados comparativos de la precisión integral de varios algoritmos. en 11 entornos desafiantes dentro del conjunto de datos OTB-100. De la Tabla 4 se desprende que el algoritmo ARCF-ICO ocupa el primer lugar en precisión en los 11 entornos, con el Los algoritmos AutoTrack y LDES ocupan el segundo lugar en cinco y tres entornos, respectivamente. Además, los algoritmos MCCT-H y Staple logran el segundo lugar en el Fast Motion y desafíos de rotación fuera del plano, respectivamente. Por lo tanto, la comparación resulta de La Tabla 4 valida aún más la efectividad del algoritmo ARCF-ICO.

Tabla 4. Tabla de comparación de precisión de OTB-100 para varios escenarios de desafío del conjunto de datos.

Nombre del desafío	Nuestro algoritmo	Algoritmo de contraste					
	ARCF-ICO	AutoTrack	LDES	MCCT-H	fDSST	Grapa	KCF
FM	0.517	0,485	0,497	0.505	0.422	0.388	0.263
[illegible]	0.483	0,476	0,496	0,476	0.412	0.433	0.314
MEGABYTE	0.514	0.502	0.467	0.443	0.402	0.367	0,278
DEF	0,482	0,477	0,472	0.425	0,378	0.336	0.266
IV	0.541	0.538	0.533	0.501	0.445	0.388	0.325
[illegible]	0,477	0,432	0.375	0.468	0.305	0.325	0.213
LR	0.533	0.532	0.596	0,492	0.462	0,439	0.306
OCC	0,488	0,482	0,457	0,426	0.430	0.422	0,239
OPR	0.563	0.552	0.512	0.423	0,458	0.563	0.325
VO	0.534	0.529	0.505	0.412	0.433	0.368	0.336
SV	0.545	0.537	0.546	0,447	0,448	0,392	0.268

5.5. Visualización de simulación

La Figura 5 ilustra los resultados de asignación de tareas obtenidos por el algoritmo ARCF-ICO. demostrando que este problema de asignación dinámica es esencialmente una programación no lineal problema con soluciones óptimas. El panel izquierdo de la Figura 5 muestra las soluciones de Pareto. para la asignación de tareas de UAV y la planificación de trayectorias derivadas del conjunto de datos UAV123@10fps, mientras que el panel derecho presenta las soluciones del conjunto de datos OTB-100. El frente de Pareto en Ambos paneles indican las compensaciones entre diferentes objetivos, mostrando la eficiencia. y eficacia del algoritmo ARCF-ICO en el manejo de la optimización multiobjetivo. En Además, también mostramos los resultados de la asignación de tareas de simulación de MATLAB de 10 UAV en

Figura 6. Como puede verse en la Figura 6d, 10 UAV están a punto de encontrar el objeto objetivo correspondiente.

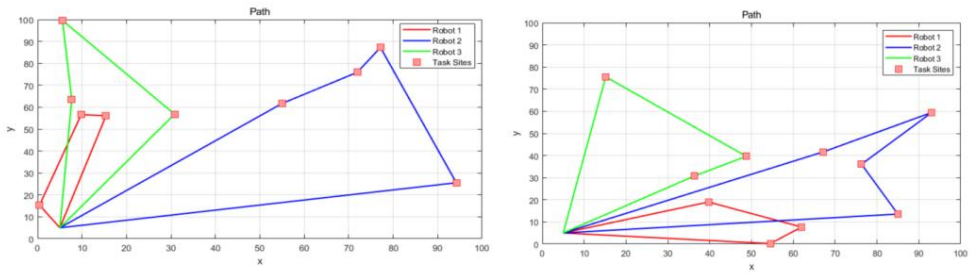


Figura 5. Asignación de tareas de solución del algoritmo ARCF-ICO. Soluciones de Pareto para la asignación de tareas y planificación de trayectorias de UAV derivadas del conjunto de datos UAV123@10fps (izquierda) y el conjunto de datos OTB-100 (derecha).

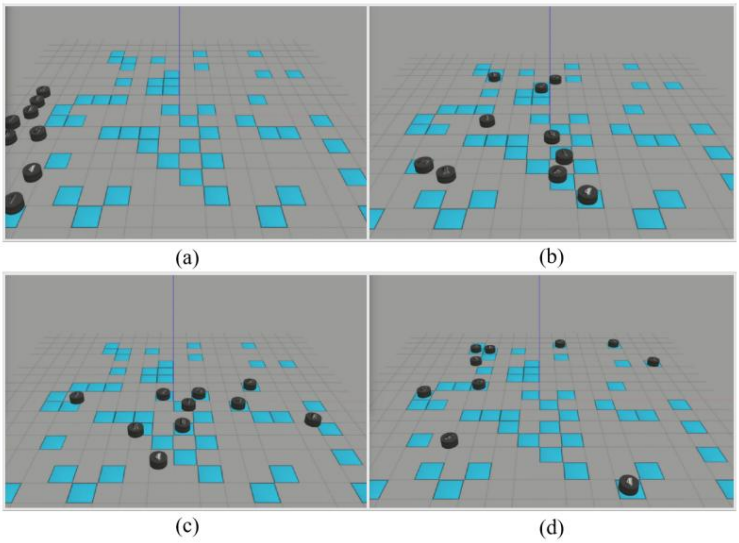


Figura 6. Demostración de asignación de tareas de simulación MATLAB de 10 UAV. (a) El estado inicial del UAV; (b,c) Estado intermedio de la asignación de tareas del UAV; (d) Estado final de la asignación de UAV.

Además, se realizan pruebas de simulación en escenas del conjunto de datos OTB-100. El mapa ambiental está dividido en cinco submapas, pero el uso exclusivo del algoritmo ARCF-ICO para la planificación de rutas de UAV puede no lograr la velocidad de entrenamiento del modelo más rápida. Esto podría deberse al proceso continuo de aprendizaje de prueba y error del algoritmo original ARCF en el entorno, que requiere más tiempo. La Figura 7 ilustra la planificación de trayectoria de las tareas de UAV en los mapas ambientales recortados del conjunto de datos OTB-100 utilizando el algoritmo ARCF-ICO.

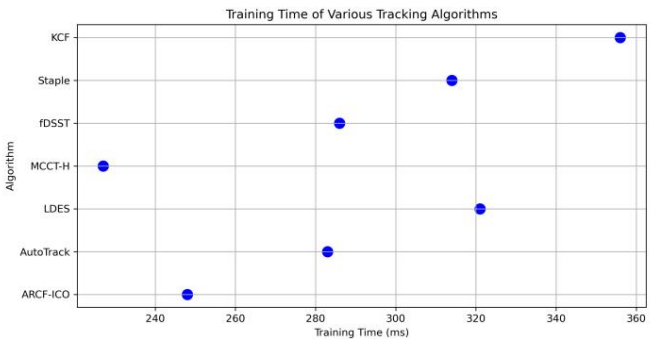


Figura 7. El algoritmo ARCF-ICO se demuestra en el mapa del entorno de cultivo del conjunto de datos OTB-100 para la planificación de misiones de UAV. El punto azul representa el tiempo de entrenamiento de cada algoritmo.

6. Conclusiones y discusión

Este estudio presenta el algoritmo ARCF-ICO para abordar la optimización multiobjetivo en la planificación de misiones de vehículos aéreos no tripulados, mejorando el rendimiento del seguimiento de vehículos aéreos no tripulados en entornos complejos. Al centrarse tanto en la convergencia como en la diversidad, el algoritmo ARCF-ICO se adapta a los rápidos cambios ambientales y las demandas dinámicas de la misión a través de actualizaciones de filtros en tiempo real. La validación utilizando los conjuntos de datos UAV123@10fps y OTB-100 demuestra que el algoritmo ARCF-ICO supera a los métodos existentes en AUC y métricas de precisión, lo que indica solidez y precisión de seguimiento superiores. Sin embargo, se identifican dos limitaciones principales: el rendimiento subóptimo del algoritmo con objetivos dinámicos de velocidad extremadamente alta y la necesidad de mejorar la eficiencia computacional. Las investigaciones futuras explorarán estructuras algorítmicas más eficientes para reducir el consumo de recursos computacionales y al mismo tiempo optimizar la respuesta a objetivos en movimiento a alta velocidad. En general, el algoritmo ARCF-ICO ofrece un avance tecnológico significativo para la optimización multiobjetivo en vehículos aéreos no tripulados, proporcionando un valor teórico y práctico sustancial para aplicaciones civiles y

Contribuciones del autor: Conceptualización, metodología, redacción: preparación del borrador original, visualización y supervisión: ZZ (Zijie Zheng). Validación, análisis formal e investigación: ZZ (Zhijun Zhang). Software, validación, redacción: revisión y edición: ZL Recursos, curación de datos, redacción: revisión y edición: QY Validación, análisis formal y metodología: YJ Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: Esta investigación no recibió financiamiento externo.

Declaración de disponibilidad de datos: Dirección de descarga del conjunto de datos del UAV123: https://cemse.kaust.edu.sa/ivul/ uav123 (consultado el 10 de marzo de 2024). Dirección de descarga del conjunto de datos OTB: http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html (consultado el 10 de marzo de 2024). El código que respalda los hallazgos de este estudio no está disponible públicamente, pero puede obtenerse del autor correspondiente previa solicitud razonable.

Conflictos de intereses: Los autores declaran no tener conflictos de intereses.

Referencias

1. Chen, J.; Zhang, Y.; Wu, L.; usted, T.; Ning, X. Un algoritmo adaptativo basado en agrupamiento para la planificación automática de rutas de heterogéneos UAV. Traducción IEEE. Intel. Transp. Sistema. 2021, 23, 16842–16853. [Referencia cruzada]
2. Chen, J.; Ling, F.; Zhang, Y.; usted, T.; Liu, Y.; Du, X. Planificación de rutas de cobertura de vehículos aéreos no tripulados heterogéneos basada en sistema de colonias de hormigas. Evolución del enjambre. Computadora. 2022, 69, 101005. [Referencia cruzada]
3. Chen, J.; Iluminado.; Zhang, Y.; usted, T.; Lu, Y.; Tiwari, P.; Kumar, N. Aprendizaje de refuerzo basado en la atención global y local para el control del comportamiento cooperativo de múltiples vehículos aéreos no tripulados. Traducción IEEE. Veh. Tecnología. 2024, 73, 4194–4206. [Referencia cruzada]
4. Ning, X.; Tian, W.; Él, F.; Bai, X.; Sol, L.; Li, W. Modelo neuronal de función de cobertura de hipersalchicha y algoritmo de aprendizaje para clasificación de imágenes. Reconocimiento de patrones. 2023, 136, 109216. [Referencia cruzada]
5. Yao, Y.; Liu, Z. El nuevo concepto de desarrollo ayuda a acelerar la formación de una nueva productividad de calidad: lógica teórica y rutas de implementación. Universidad J. Xi'an. Finanzas. Economía. 2024, 37, 3–14.
6. Mamá, G.; Chen, X. De potencia financiera a potencia financiera: comparación internacional y enfoque de China. Universidad J. Xi'an. Finanzas. Economía. 2024, 37, 46–59.
7. Wang, J.; Li, F.; Cualquier.; Zhang, X.; Sun, H. Hacia una fusión robusta de cámaras LiDAR en el espacio BEV mediante atención mutua deformable y Agregación Temporal. Traducción IEEE. Sistema de circuitos. Tecnología de vídeo. 2024, 34, 5753–5764. [Referencia cruzada]
8. Chen, CC; Wei, CC; Chen, SH; Sol, LM; Lin, HH AI predijo el modelo de competencia para maximizar el desempeño laboral. Cibern. Sistema. 2022, 53, 298–317. [Referencia cruzada]
9. Kaufmann, E.; Bauersfeld, L.; Loquercio, A.; Müller, M.; Koltún, V.; Scaramuzza, D. Carreras de drones a nivel de campeón mediante aprendizaje por refuerzo profundo. Naturaleza 2023, 620, 982–987. [Referencia cruzada]
10. Uthamacumaran, A. Detección de patrones en el paisaje de Waddington del glioblastoma a través de redes generativas adversas. Cibern. Sistema. 2022, 53, 223–237. [Referencia cruzada]
11. Ning, E.; Wang, C.; Zhang, H.; Ning, X.; Tiwari, P. Reidentificación de personas ocluidas con el aprendizaje profundo: una encuesta y perspectivas. Sistema experto. Aplica. 2024, 239, 122419. [Referencia cruzada]
12. Li, P.; Duan, H. Planificación de rutas de vehículos aéreos no tripulados basada en un algoritmo de búsqueda gravitacional mejorado. Ciencia. Tecnología China. Ciencia. 2012, 55, 2712–2719. [Referencia cruzada]
13. Qu, C.; Gai, W.; Zhang, J.; Zhong, M. Un novedoso algoritmo optimizador híbrido de lobo gris para la planificación de rutas de vehículos aéreos no tripulados (UAV). Sistema basado en el conocimiento. 2020, 194, 105530. [Referencia cruzada]

14. Dasdemir, E.; Köksalan, M.; Öztürk, DT Un algoritmo evolutivo multiobjetivo basado en puntos de referencia flexible: una aplicación al problema de planificación de rutas de vehículos aéreos no tripulados. Computadora. Ópera. Res. 2020, 114, 104811. [Referencia cruzada]
15. Yao, P.; Wang, H.; Ji, H. Objetivo de seguimiento de múltiples UAV en entornos urbanos mediante control predictivo de modelo y optimizador de lobo gris mejorado. Aerosp. Ciencia. Tecnología. 2016, 55, 131-143. [Referencia cruzada]
16. Papaioannou, S.; Laoudias, C.; Kolios, P.; Teocarides, T.; Panayiotou, CG Estimación y control conjunto para monitoreo pasivo de múltiples objetivos con un agente UAV autónomo. En Actas de la 31.ª Conferencia Mediterránea sobre Control y Automatización (MED) de 2023, Limassol, Chipre, 26 a 29 de junio de 2023; págs. 176–181.
17. Ren, Q.; Yao, Y.; Yang, G.; Zhou, X. Planificación de rutas multiobjetivo para UAV en el entorno urbano basada en CDNSGA-II. En Actas de la Conferencia Internacional IEEE de 2019 sobre Ingeniería de Sistemas Orientados a Servicios (SOSE), San Francisco, CA, EE. UU., 4 al 9 de abril de 2019; págs. 350–3505.
18. Precio, KV Una introducción a la evolución diferencial. En Nuevas Ideas en Optimización; McGraw: Berkshire, Reino Unido, 1999; págs. 79-108.
19. Gad, AG Algoritmo de optimización de enjambre de partículas y sus aplicaciones: una revisión sistemática. Arco. Computadora. Métodos Ing. 2022, 29, 2531–2561. [Referencia cruzada]
20. Yu, X.; Jiang, N.; Wang, X.; Li, M. Un algoritmo híbrido basado en el optimizador de lobo gris y la evolución diferencial para la planificación de rutas de vehículos aéreos no tripulados. Sistema experto. Aplica. 2023, 215, 119327. [Referencia cruzada]
21. Liu, J.; Wei, X.; Huang, H. Un algoritmo mejorado de optimización del lobo gris y su aplicación en la planificación de rutas. Acceso IEEE 2021, 9, 121944–121956. [Referencia cruzada]
22. Xue, J.; Shen, B.; Pan, A. Un algoritmo de búsqueda de gorriones intensificado para resolver problemas de optimización. J. Ambiente. Intel. Humaniz. Computadora. 2023, 14, 9173–9189. [Referencia cruzada]
23. Sahingoz, OK Planificación de rutas de vuelo para un sistema multi-UAV con algoritmos genéticos y curvas de Bézier. En Actas de la Conferencia Internacional sobre Sistemas de Aeronaves No Tripuladas (ICUAS) de 2013, Atlanta, GA, EE. UU., 28 a 31 de mayo de 2013; págs. 41–48.
24. Cui, Z.; Zhang, J.; Wu, D.; Cai, X.; Wang, H.; Zhang, W.; Chen, J. Algoritmo híbrido de optimización de enjambre de partículas de muchos objetivos para el problema de la producción de carbón verde. Inf. Ciencia. 2020, 518, 256–271. [Referencia cruzada]
25. Zhang, J.; Xue, F.; Cai, X.; Cui, Z.; Chang, Y.; Zhang, W.; Li, W. Protección de la privacidad basada en un algoritmo de optimización de muchos objetivos. Concurrir. Computadora. Practica. Exp. 2019, 31, e5342. [Referencia cruzada]
26. Gong, D.; Sol, J.; Miao, Z. Un algoritmo genético basado en conjuntos para problemas de optimización de intervalos con muchos objetivos. Traducción IEEE. Evolución. Computadora. 2016, 22, 47–60. [Referencia cruzada]
27. Yang, S.; Li, M.; Liu, X.; Zheng, J. Un algoritmo evolutivo basado en cuadrículas para la optimización de muchos objetivos. Traducción IEEE. Evolución. Computadora. 2013, 17, 721–736. [Referencia cruzada]
28. Cui, Z.; Chang, Y.; Zhang, J.; Cai, X.; Zhang, W. NSGA-III mejorado con operador de selección y eliminación. Evolución del enjambre. Computadora. 2019, 49, 23–33. [Referencia cruzada]
29. Hobbie, JG; Gandomi, AH; Rahimi, I. Una comparación de técnicas de manejo de restricciones en NSGA-II. Arco. Computadora. Métodos Ing. 2021, 28, 3475–3490. [Referencia cruzada]
30. Liu, Y.; Gong, D.; Sol, J.; Jin, Y. Un algoritmo evolutivo de muchos objetivos que utiliza una estrategia de selección uno por uno. Traducción IEEE. Cibern. 2017, 47, 2689–2702. [Referencia cruzada] [PubMed]
31. Gao, X.; Zhang, H.; Song, S. Un algoritmo evolutivo basado en propiedades de regularidad para la optimización multiobjetivo. Evolución del enjambre. Computadora. 2023, 78, 101258. [Referencia cruzada]
32. Bao, C.; Gao, D.; Gu, W.; Xu, L.; Goodman, ED Un nuevo algoritmo evolutivo basado en descomposición adaptativa para múltiples y optimización de muchos objetivos. Sistema experto. Aplica. 2023, 213, 119080. [Referencia cruzada]
33. Liu, S.; Lin, Q.; Wong, Kansas; Coello, CAC; Li, J.; Ming, Z.; Zhang, J. Una estrategia de vector de referencia autoguiada para muchos objetivos mejoramiento. Traducción IEEE. Cibern. 2020, 52, 1164–1178. [Referencia cruzada]
34. Yi, J.; Zhang, W.; Bai, J.; Zhou, W.; Yao, L. Algoritmo evolutivo multifactorial basado en descomposición dinámica mejorada para problemas de optimización de muchos objetivos. Traducción IEEE. Evolución. Computadora. 2021, 26, 334–348. [Referencia cruzada]
35. Nondy, J.; Gogoi, TK Comparación de rendimiento de algoritmos evolutivos multiobjetivo para exergéticos y exergoambientales optimización de un sistema combinado de calor y electricidad de referencia. Energía 2021, 233, 121135. [CrossRef]
36. Cai, X.; Hu, Z.; Chen, J. Un algoritmo de recomendación de optimización de muchos objetivos basado en la minería de conocimientos. Inf. Ciencia. 2020, 537, 148–161. [Referencia cruzada]
37. Cai, T.; Wang, H. Un método de análisis de convergencia general para un algoritmo de optimización evolutivo multiobjetivo. Inf. Ciencia. 2024, 663, 120267. [Referencia cruzada]
38. Davies, L.; Bolam, RC; Vagapov, Y.; Anuchin, A. Revisión de tecnologías de sistemas de aeronaves no tripuladas para permitir operaciones más allá de la línea de visión (BVLOS). En Actas de la X Conferencia Internacional sobre Sistemas de Propulsión de Energía Eléctrica (ICEPDS) de 2018, Novocherkassk, Rusia, 3 a 6 de octubre de 2018; págs. 1–6.
39. Mueller, M.; Smith, N.; Ghanem, B. Un punto de referencia y un simulador para el seguimiento de vehículos aéreos no tripulados. En Actas de Computer Vision – ECCV 2016: 14ª Conferencia Europea, Ámsterdam, Países Bajos, 11 a 14 de octubre de 2016; Actas, Parte I 14; Springer: Berlín/Heidelberg, Alemania, 2016; págs. 445–461.
40. Wu, Y.; Lim, J.; Yang, MH Seguimiento de objetos en línea: un punto de referencia. En actas de la Conferencia IEEE sobre visión por computadora y Reconocimiento de patrones, Portland, Oregón, EE. UU., 23 a 28 de junio de 2013; págs. 2411–2418.
41. Henriques, JF; Caseiro, R.; Martín, P.; Batista, J. Seguimiento de alta velocidad con filtros de correlación kernelizados. Traducción IEEE. Patrón Anal. Mach. Intel. 2015, 37, 583–596. [Referencia cruzada]

42. Li, Y.; Zhu, J.; Hoi, Carolina del Sur; Canción, W.; Wang, Z.; Liu, H. Estimación robusta de la transformación de similitud para el seguimiento visual de objetos. En Actas de la Conferencia AAAI sobre Inteligencia Artificial, Honolulu, Hawaii, EE. UU., 27 de enero a 1 de febrero de 2019; Volumen 33, págs. 8666–8673.
43. Wang, N.; Zhou, W.; Tian, Q.; Hong, R.; Wang, M.; Li, H. Filtros de correlación de señales múltiples para un seguimiento visual sólido. En Actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones, Salt Lake City, UT, EE. UU., 18 a 23 de junio de 2018; págs. 4844–4853.
44. Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; Torr, PH Staple: Aprendices complementarios para seguimiento en tiempo real. En Actas de la Conferencia IEEE sobre visión por computadora y reconocimiento de patrones, Las Vegas, NV, EE. UU., 27 a 30 de junio de 2016; págs. 1401-1409.
45. Danelljan, M.; Häger, G.; Khan, FS; Felsberg, M. Seguimiento espacial a escala discriminativa. Traducción IEEE. Patrón Anal. Mach. Intel. 2016, 39, 1561-1575. [Referencia cruzada]
46. Li, Y.; Fu, C.; Ding, F.; Huang, Z.; Lu, G. AutoTrack: Hacia el seguimiento visual de alto rendimiento para UAV con regularización espacio-temporal automática. En Actas de la Conferencia IEEE/CVF sobre visión por computadora y reconocimiento de patrones, Seattle, WA, EE. UU., 17 a 21 de junio de 2020; págs. 11923-11932.

electrónica

MDPI

Artículo

Optimización de la ubicación de los sensores y técnicas de aprendizaje automático para una clasificación precisa de los gestos de las manos

Lakshya Chaplot [1], Sara Houshmand, [1] Karla Beltrán Martínez [1], John Andersen 2,3 y Hossein Rouhani 1,3,*

[1] Departamento de Ingeniería Mecánica, Universidad de Alberta, Edmonton, AB T6G 2R3, Canadá; lakshyachaplot@gmail.com (LC); shoushma@ualberta.ca (SH); beltranm@ualberta.ca (KBM)
[2] Departamento de Pediatría, Universidad de Alberta, Edmonton, AB T6G 2R3, Canadá;
john.andersen@albertahealthservices.ca 3 Glenrose Rehabilitation Hospital, Edmonton, AB T5G 0B7, Canadá * Correspondencia: hrouhani@ualberta.ca

Resumen: Millones de personas viven con amputaciones de extremidades superiores, lo que las convierte en beneficiarios potenciales de prótesis de manos y brazos. Si bien las prótesis mioeléctricas han evolucionado para satisfacer las necesidades de los amputados, los desafíos siguen relacionados con su control. Esta investigación aprovecha sensores de electromiografía de superficie y técnicas de aprendizaje automático para clasificar cinco gestos fundamentales de las manos. Al utilizar características extraídas de datos de electromiografía, empleamos un clasificador de máquina de vectores de soporte no lineal, basado en aprendizaje de múltiples núcleos, para el reconocimiento de gestos. Nuestro conjunto de datos abarcó a ocho participantes jóvenes sin discapacidades. Además, nuestro estudio realizó un análisis comparativo de cinco configuraciones distintas de ubicación de sensores. Estas configuraciones capturan datos de electromiografía asociados con los movimientos del dedo índice y el pulgar, así como con los movimientos del dedo índice y anular . También comparamos cuatro clasificadores diferentes para determinar cuál es el más capaz de clasificar los gestos con las manos. La configuración de doble sensor colocada estratégicamente para capturar los movimientos del pulgar y el índice fue la más efectiva: esta configuración de doble sensor logró una precisión del 90 % para clasificar los cinco gestos utilizando el clasificador de máquina de vectores de soporte. Además, la aplicación del aprendizaje de múltiples núcleos dentro del clasificador de máquina de vectores de soporte muestra su eficacia, logrando la mayor precisión de clasificación entre todos los clasificadores. Este estudio mostró el potencial de los sensores de electromiografía de superficie y el aprendizaje automático para mejorar el control y la funcionalidad de las prótesis mioeléctricas para personas con amputaciones de extremidades superiores.

Palabras clave: sensor mioeléctrico; gesto manual; máquinas de vectores soporte; mano protésica; clasificación; aprendizaje automático

Cita: Chaplot, L.; Houshmand, S.; Martínez, KB; Andersen, J.; Rouhani, H. Optimización de la ubicación de sensores y técnicas de aprendizaje automático para Clasificación precisa de gestos con las manos. Electrónica 2024, 13 , 3072. https://doi.org/10.3390/electronics13153072

Editor académico: Janos Botzheim

Recibido: 7 de junio de 2024
Revisado: 24 de julio de 2024
Aceptado: 1 de agosto de 2024
Publicado: 3 de agosto de 2024

1. Introducción

En 2017, el recuento mundial de amputaciones unilaterales de miembros superiores superó los 11,3 millones, y 11,0 millones de personas adicionales experimentaron amputaciones bilaterales de miembros superiores [1]. En Canadá, alrededor de 6.800 personas viven con una amputación proximal a la muñeca [2]. Un estudio reciente [2] comparó los resultados de utilidad y los costos asociados con dos intervenciones para el tratamiento de amputaciones de manos: alotrasplante compuesto vascularizado de mano y prótesis de mano mioeléctricas. La conclusión fue que tratar las amputaciones unilaterales con prótesis mioeléctricas era más rentable.

Las prótesis de manos mioeléctricas han surgido como una vía fundamental para restaurar tanto el gesto como las capacidades prensiles en amputados de miembros superiores, ofreciendo una alternativa no invasiva a las intervenciones quirúrgicas permanentes [2]. Los sistemas de control predominantes en las prótesis a menudo utilizan un mecanismo basado en un disparador que se basa en uno o dos canales de electromiografía de superficie (EMG) [3,4]. Esta configuración asigna eventos de contracción de un solo músculo a secuencias de movimiento predefinidas, lo que requiere comandos explícitos del usuario para cambiar de modo [3,4]. Este cambio secuencial introduce latencia en los tiempos de ejecución, lo que requiere múltiples comandos distintos para realizar la transición entre diferentes modos de agarre [3,4]. La naturaleza r

Este proceso de cambio de agarre, junto con una dinámica de control incómoda y una falta de retroalimentación suficiente, se ha identificado como el principal contribuyente a las bajas tasas de aceptación observadas para los dispositivos de prótesis mioeléctricas [5,6]. Abordar estos desafíos es esencial para mejorar la usabilidad y aceptación de los dispositivos protésicos mioeléctricos dentro de la comunidad de usuarios [5,6].

En trabajos anteriores se han documentado varios esfuerzos para clasificar las señales sEMG (electromiografía de superficie) de los músculos del antebrazo humano . Para mitigar las preocupaciones sobre la intuición, las soluciones prototipo dentro de la literatura existente se centran en descifrar las intenciones de los gestos del usuario apuntando a distintos músculos flexores y extensores en el antebrazo [7]. Las contracciones voluntarias de los músculos restantes del antebrazo después de la amputación se pueden identificar mediante clasificadores de aprendizaje automático, como clasificadores de redes neuronales artificiales (ANN), análisis discriminante lineal (LDA) y máquinas de vectores de soporte (SVM). A menudo se elige SVM debido a su interpretabilidad matemática y optimización global. Funciona bien incluso con un conjunto de entrenamiento pequeño [8]. El parámetro de elasticidad de SVM, también conocido como hiperparámetro C de restricción de caja, controla la penalización máxima impuesta a las observaciones que violan el margen y ayuda a prevenir el sobreajuste [8]. Palkowski y Redlarski [9] emplearon dos sensores EMG en el antebrazo, tomando datos a 16 Hz, para discernir seis gestos completos de la mano y la muñeca utilizando un clasificador SVM. Lee y cols. [10] clasificaron con éxito diez gestos con las manos utilizando características obtenidas por tres sensores EMG y lograron una precisión superior al 90% para cada participante. Sin embargo, su modelo de aprendizaje automático se sometió a entrenamiento y pruebas en conjuntos de datos de participantes sin fusión de datos entre participantes para evaluar la capacidad de generalización del modelo. Este enfoque de prueba restrictivo introduce un sesgo en la precisión de la clasificación, lo que plantea dudas sobre la aplicabilidad de esta tecnología en prótesis de manos. Otros trabajos anteriores [8,11,12] lograron una alta precisión de más del 90% para clasificar múltiples gestos completos de la mano y la muñeca, como apertura/cierre de la muñeca, desviación cubital y radial, y flexión-extensión. Sin embargo, los gestos completos de manos y muñecas son menos difíciles de clasificar y ofrecen una aplicación funcional limitada para los amputados de miembros superiores que buscan restaurar la destreza manual.

La eficacia de apuntar a músculos específicos requiere el despliegue estratégico de electrodos y su configuración, una consideración crítica para los enfoques basados en características que están relativamente inexplorados en la literatura.

Este proyecto tuvo como objetivo mejorar la precisión de la clasificación de los gestos de las manos, la ubicación estratégica de los sensores y la selección de gestos prácticos para una posible integración con prótesis de manos mioeléctricas. Por lo tanto, este estudio investigó el desarrollo de un clasificador SVM basado en aprendizaje de núcleo múltiple (MKL) para clasificar cinco gestos manuales complejos y cruciales para los amputados: agarre con fuerza (apretar la muñeca), agarre con gancho (agarre de cuatro dígitos), pellizco fino (usando el dedo índice y el pulgar), pellizco grueso (usando los cinco dedos) y gesto de señalar (flexión de los dedos 3, 4 y 5). La metodología de investigación implicó construir el clasificador utilizando datos recopilados a través de cinco configuraciones distintas de sensores, cada una utilizando uno o dos sensores EMG en el antebrazo hacia un enfoque de recopilación de datos minimalista, mientras se esforzaba por identificar la ubicación óptima del sensor para obtener la mayor precisión de clasificación.

2. Materiales y métodos

2.1. Procedimiento

experimental Se recogieron datos de sEMG de la mano derecha de ocho participantes sin deterioro ni diagnóstico neurológico/musculoesquelético (edad: 21 ± 2 años (media ± DE), altura corporal: 169,5 ± 2,8 cm (media ± DE), peso corporal : 57,9 ± 9,5 kg (media ± DE), 7 hombres y 1 mujer). Todos los participantes estaban familiarizados con los procedimientos experimentales. Se obtuvo el consentimiento informado de todos los sujetos involucrados en el estudio. El estudio se realizó de acuerdo con la Declaración de Helsinki y fue aprobado por el Comité de Ética Institucional de la Universidad de Alberta (AB T6G 2N2, aprobado el 25 de abril de 2022).

Las señales de sEMG se adquirieron utilizando sensores musculares bipolares MyoWare 2.0 [13] (SparkFun Electronics, Niwot, CO, EE. UU.) elegidos por su potencial de integración en prótesis de mano de bajo costo. La versión anterior de este sensor se ha utilizado frecuentemente en

Las señales de sEMG se adquirieron utilizando sensores musculares bipolares MyoWare 2.0 [13] (SparkFun Electronics, Niwot, CO, EE. UU.) elegidos por su potencial de integración en prótesis de mano de bajo costo. La versión anterior de este sensor se ha utilizado con frecuencia en la literatura debido a su bajo costo, características fáciles de personalizar e informes de rendimiento favorables en estudios de validación, lo que demuestra que es comparable a modelos más costosos . su bajo costo, características fáciles de personalizar y rendimiento favorable sistemas EMG comerciales [14,15]. MyoWare 2.0 tiene tres electrodos (en la mitad del músculo y en el extremo del músculo), según informes de estudios de validación que demuestran que es comparable a los electrodos comerciales más caros. cie y referencia [13]. Antes de ser adquirida por el microcontrolador, la señal diferencial

Sistemas EMG [14,15]. MyoWare 2.0 tiene tres electrodos: medio muscular, final muscular y paso a través de un amplificador de instrumentación con un CMRR alto (referencia de rechazo de modo común [13]. Antes de ser adquirida por el microcontrolador, la señal diferencial pasó relación, que es la relación entre la ganancia diferencial y la ganancia en modo común de la etapa del amplificador) a través de un amplificador de instrumentación con un alto CMRR (relación de rechazo de modo común (140 dB) y ganancia unitaria para eliminar fuentes de ruido comunes, como el ruido de línea de 50 Hz. que es la relación entre la ganancia diferencial y la ganancia en modo común de la etapa del amplificador) (140 dB) Posteriormente, la señal fue filtrada por un filtro de paso de banda de primer orden con frecuencia de corte y ganancia unitaria para eliminar fuentes de ruido comunes, como el ruido de línea de 50 Hz. Subsecciones a 20 Hz y 498 Hz [13]. Después de esto, la señal se sometió a rectificación y, posteriormente, la señal se filtró mediante un filtro de paso de banda de primer orden con frecuencias de corte en

suavizado, logrado mediante un circuito de detección de envolvente de paso bajo (3,6 Hz) integrado en los 20 Hz y 498 Hz [13]. A continuación, la señal se sometió a rectificación y suavizado, hardware del sensor, lo que resulta en una envolvente lineal de la señal EMG. La EMG envolvente lograda por un circuito de detección de envolvente de paso bajo (3,6 Hz) integrado en la salida dura del sensor fue luego adquirida por el microcontrolador (Arduino UNO) con un software de muestreo , lo que dio como resultado una envolvente lineal de la señal EMG. La salida EMG envuelta fue frecuencia de 780 Hz (SD = 5 Hz) para abordar las inquietudes relacionadas con el submuestreo previo adquirido por el microcontrolador (Arduino UNO) con una frecuencia de muestreo de 780 Hz estudios [16]. (SD = 5 Hz) para abordar las preocupaciones relacionadas con el submuestreo en estudios anteriores [16].

Las configuraciones de ubicación del sensor EMG se basaron en los gestos destinados a ser Las configuraciones de ubicación del sensor EMG se basaron en los gestos destinados a ser clasificado. Los sensores se colocaron sólo a lo largo de los músculos del antebrazo clasificados como los más responsables. Los sensores se colocaron sólo a lo largo de los músculos del antebrazo que son los principales responsables para la flexión del dedo índice, anular y pulgar. Los cinco gestos utilizados fueron (1) a para la flexión del dedo índice, anular y pulgar. Los cinco gestos utilizados fueron (1) un pellizco grueso usando los cinco dedos, (2) un pellizco fino usando el dedo índice y el pulgar, (3) un gancho gancho usando los cinco dedos, (2) un pellizco fino usando el dedo índice y el pulgar, (3) un gancho agarre (agarre de cuatro dígitos), (4) gesto de señalar (flexión de los dígitos 3, 4 y 5) y (5) agarre de fuerza (agarre de cuatro dígitos), (4) gesto de señalar (flexión de los dígitos 3, 4 y 5), y (5) una comprensión del poder. agarre (apretar la muñeca), etiquetados como 0 a 4, respectivamente, para el clasificador. Estos cinco (apretar la muñeca), etiquetados como 0 a 4, respectivamente, para el clasificador. Estos cinco gestos

Se considera que los gestos (Figura 1) tienen una gran utilidad para los amputados en su vida diaria y (Figura 1) tienen una gran utilidad para los amputados en su vida diaria y también son También se observan en prótesis comerciales [3,4]. visto en prótesis comerciales [3,4].

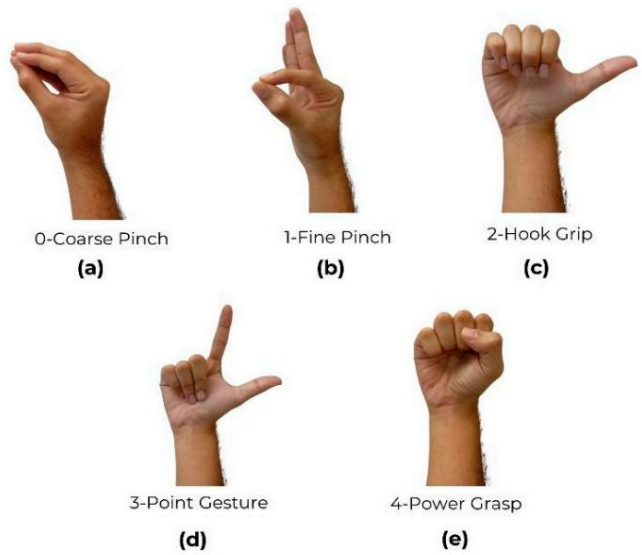


Figura 1. Los cinco gestos diferentes con sus correspondientes etiquetas de predicción: (a) pellizco grueso, (b) (b) pellizco fino, (c) agarre con gancho, (d) gesto señalador y (e) agarre de fuerza. Figura 1. Los cinco gestos diferentes con sus correspondientes etiquetas de predicción: (a) pellizco grueso, pellizco fino, (c) agarre con gancho, (d) gesto señalador y (e) agarre de fuerza.

Los sensores EMG se colocaron en tres ubicaciones en el lado ventral del antebrazo. Los sensores EMG se colocaron en tres ubicaciones en el lado ventral del antebrazo (Figura 2), para diferentes configuraciones para adquirir señales EMG de los músculos principalmente responsables de la flexión del dedo índice, anular y pulgar. el sensor. Principalmente responsable de la flexión del dedo índice, anular y pulgar. Las configuraciones del sensor se basaron en [17–19] y se describen a continuación: Las figuraciones se basaron en [17-19] y se describen a continuación: C1: un sensor

- colocado proximal a la muñeca a lo largo del flexor largo del pulgar para adquirir de flexión del pulgar.
- C1: un sensor colocado proximal a la muñeca a lo largo del flexor largo del pulgar para adquirir los datos los datos de flexión del
- pulgar. • C2: un sensor colocado proximal a la muñeca a lo largo del flexor superficial de los dedos para C2: un sensor colocado proximal a la muñeca a lo largo del flexor superficial de los dedos para adquirir los datos de flexión del dedo índice. adquirir los datos de flexión del dedo índice. • C3:
- un sensor colocado a lo largo del flexor superficial de los dedos y el flexor de los dedos profundus para adquirir datos de flexión del dedo anular.
- C4: Dos sensores colocados proximales a la muñeca a lo largo del flexor superficial de los dedos para adquirir datos de flexión del dedo índice y proximal al codo a lo largo del flexor de los dedos superficial y flexor profundo de los dedos para adquirir datos de flexión del dedo anular.

- C3: Un sensor colocado a lo largo del flexor superficial de los dedos y el flexor superficial de los dedos.
  fondo de ojo para adquirir datos de flexión del dedo anular.
- C4: Dos sensores colocados proximalmente a la muñeca a lo largo del flexor superficial de los dedos para adquirir
  datos de flexión del dedo índice y proximales al codo a lo largo del flexor superficial de los dedos y el flexor
  profundo de los dedos para adquirir datos de flexión del dedo anular. • C5: Dos sensores colocados
  proximalmente a la muñeca a lo largo del flexor superficial de los dedos C5: Dos sensores colocados proximalmente a
  la muñeca a lo largo del flexor superficial de los dedos
  para adquirir datos de flexión del dedo índice y a lo largo del flexor largo del pulgar para adquirir datos de flexión
  del dedo índice y a lo largo del flexor largo del pulgar para adquirir el
  datos de flexión del pulgar.
  datos de flexión del pulgar.

Se pidió a los participantes que realizaran cada gesto repetidamente 20 veces, con un tiempo.
Intervalo de dos segundos entre cada repetición. Después de realizar una ronda de col- intervalo de dos segundos
selección para un solo gesto, el participante tomó un descanso de 4 minutos para relajar sus músculos antes de la
desahogar la fatiga antes de comenzar la siguiente ronda para un nuevo gesto. Las señales medidas previenen la fatiga
antes de comenzar la siguiente ronda para un nuevo gesto. Las señales medidas
EMG se registraron simultáneamente. de los dos sensores

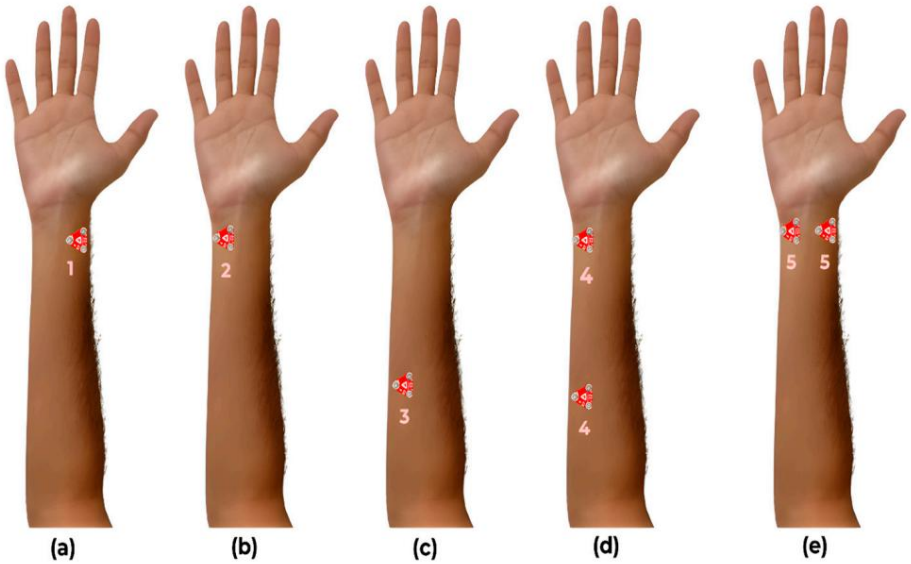


Figura 2. Las cinco configuraciones diferentes de ubicación del sensor utilizadas para adquirir datos EMG: (a) pulgar Figura 2. Las cinco configuraciones diferentes de ubicación del sensor utilizadas para adquirir datos EMG: (a) pulgar
datos de flexión del dedo índice, (c) datos de flexión del dedo índice, (c) datos de flexión del dedo anular, (d) datos de flexión del dedo
índice y anular, datos de flexión, (b) datos de flexión del dedo índice, (c) datos de flexión del dedo anular, (d) datos de índice y flexión del anillo,
y (e) datos de flexión del índice y del pulgar. y (e)
datos de flexión del índice y del pulgar.

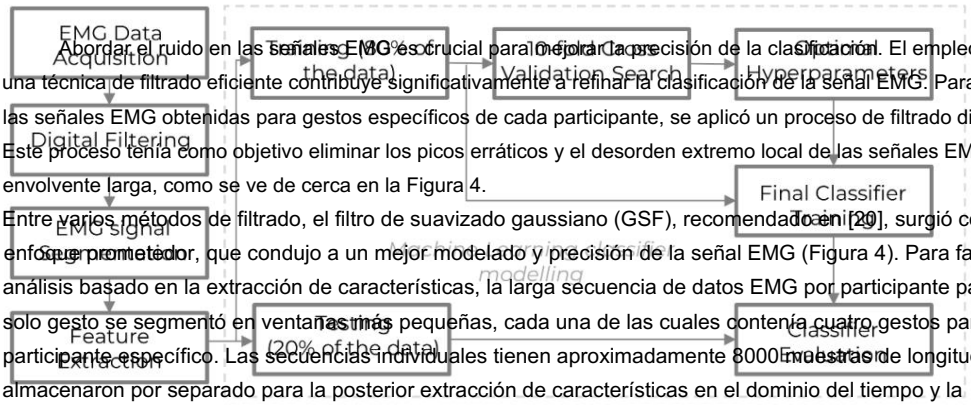
2.2 Procesamiento de datos
2.2. Procesamiento de datos

La Figura 3 muestra un diagrama de flujo de los pasos de procesamiento de
datos involucrados en todo el experimento y el proceso de modelado. Primero,
los datos de EMG fueron registrados por sensores y filtrados y modelados. Primero,
los datos de EMG fueron registrados por sensores y filtrados para
para eliminar el ruido. En segundo lugar, las señales se segmentaron en secuencias más cortas y el tiempo eliminó el
ruido. En segundo lugar, las señales se segmentaron en secuencias más cortas, y el tiempo y
y las características del dominio de la frecuencia se extrajeron para capturar características distintas del EMG.
Datos EMG. En tercer lugar, los datos procesados se utilizaron para la clasificación de gestos con las manos utilizando
ma- data. En tercer lugar, los datos procesados se utilizaron para clasificar los gestos de las manos utilizando máquinas.
aprendizaje, donde se realizó una división de 80 a 20 pruebas de entrenamiento para entrenar y evaluar los modelos.
aprendizaje chino, donde se realizó una división de 80 a 20 pruebas de entrenamiento para entrenar y evaluar el
los hiperparámetros de los modelos de aprendizaje automático para el conjunto de datos dado fueron
modelos. Cuarto, los hiperparámetros de los modelos de aprendizaje automático para el conjunto de datos dado fueron
ajustado utilizando un método de validación cruzada de búsqueda de cuadrícula de 10 veces para mejorar el rendimiento. Quinto,
se ajustaron utilizando un método de validación cruzada de búsqueda de cuadrícula de 10 veces para mejorar el rendimiento.
hiperparámetros, optimalos. En sexto lugar, el rendimiento del modelo se evaluó utilizando datos de prueba invisibles y estos
Finalmente, los clasifica- En sexto lugar, el rendimiento del modelo se evaluó utilizando datos de prueba invisibles y estos
y parámetros. En sexto lugar, el desempeño del modelo se evaluó utilizando datos de prueba invisibles y la métrica de
evaluación (la puntuación de precisión) para reflejar la efectividad del modelo en la clasificación.
los datos estáticos de EMG.
los datos EMG.

Abordar el ruido en las señales EMG es crucial para mejorar la precisión de la clasificación. El empleo
de una técnica de filtrado eficiente contribuye significativamente a refinar la señal EMG.
clasificación. Para refinar las señales EMG obtenidas para gestos específicos de cada participante , se aplicó
un proceso de filtrado digital. Este proceso tenía como objetivo eliminar picos erráticos.
y el desorden extremo local de las señales EMG de envolvente larga, como se ve de cerca en la Figura 4.
Entre varios métodos de filtrado, se recomienda el filtro de suavizado gaussiano (GSF).
en [20], surgió como un enfoque prometedor, que condujo a un mejor modelado de señales EMG.
y precisión (Figura 4). Para facilitar el análisis basado en la extracción de características, los datos largos de EMG
La secuencia por participante para un solo gesto se segmentó en ventanas más pequeñas, cada una
que contiene cuatro gestos para un participante específico. Las secuencias individuales son aproximadamente

8000 muestras de largo y se almacenaron por separado para el dominio de tiempo y frecuencia posterior.
extracción de características.

Figura 3. Diagrama de flujo de procesamiento de datos de EMG.



Abordar el ruido en las señales EMG es crucial para mejorar la precisión de la clasificación. El empleo de una técnica de filtrado eficiente contribuye significativamente a refinar la clasificación de la señal EMG. Para refinar las señales EMG obtenidas para gestos específicos de cada participante, se aplicó un proceso de filtrado digital. Este proceso tenía como objetivo eliminar los picos erráticos y el desorden extremo local de las señales EMG de envolvente larga, como se ve de cerca en la Figura 4.

Entre varios métodos de filtrado, el filtro de suavizado gaussiano (GSF), recomendado en [20], surgió como un enfoque prometedor, que condujo a un mejor modelado y precisión de la señal EMG (Figura 4). Para facilitar el análisis basado en la extracción de características, la larga secuencia de datos EMG por participante para un solo gesto se segmentó en ventanas más pequeñas, cada una de las cuales contenía cuatro gestos para un participante específico. Las secuencias individuales tienen aproximadamente 8000 muestras de longitud y se almacenaron por separado para la posterior extracción de características en el dominio del tiempo y la frecuencia.

Figura 3. Diagrama de flujo de procesamiento de datos de EMG.

Abordar el ruido en las señales EMG es crucial para mejorar la precisión de la clasificación. El empleo de una técnica de filtrado eficiente contribuye significativamente a refinar la clasificación de la señal EMG. Para refinar las señales EMG obtenidas para gestos específicos de cada participante, se aplicó un proceso de filtrado digital. Este proceso tenía como objetivo eliminar los picos erráticos y el desorden extremo local de las señales EMG de envolvente larga, como se ve de cerca en la Figura 4.

Entre varios métodos de filtrado, el filtro de suavizado gaussiano (GSF), recomendado en [20], surgió como un enfoque prometedor, que condujo a un mejor modelado y precisión de la señal EMG (Figura 4). Para facilitar el análisis basado en la extracción de características, la larga secuencia de datos EMG por participante para un solo gesto se segmentó en ventanas más pequeñas, cada una de las cuales contenía cuatro gestos para un participante específico. Las secuencias individuales tienen aproximadamente 8000 muestras de longitud y se almacenaron por separado para la posterior extracción de características en el dominio del tiempo y la frecuencia.

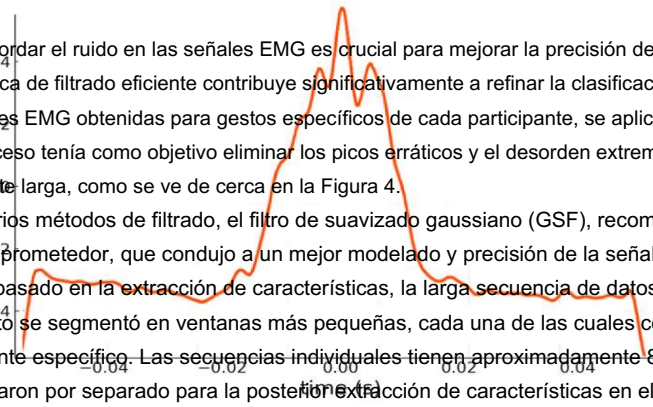


Figura 4. Señal EMG filtrada para el gesto de agarre con fuerza (se muestra un ejemplo para la configuración C3 y las transformaciones se basaron en la ondícula Daubechies-2. Se han utilizado las características anteriores

Figura 4. Señal EMG filtrada para el gesto de agarre con fuerza (se muestra un ejemplo para la configuración C3 y el gesto de captar el poder). el gesto

2.3. Extracción y selección de características

2.3. Extracción y selección de características

La extracción de características es crucial para reducir la dimensionalidad de los datos que representan un gesto, permitiendo así la clasificación de gestos. Un total de 22 características en el dominio del tiempo y la frecuencia de las 8000 muestras de señales EMG segmentadas largas para cada gesto. Las características incluían medidas estadísticas como la raíz cuadrática media (RMS), varianza (VAR), valor absoluto medio (MAV), cambio de signo de pendiente (SSC), cruces por cero (ZC), longitud de la forma de onda (WL), desviación absoluta mediana (MAD), asimetría, curtosis y energía. Estas características están bien establecidas en la literatura por su utilidad en el reconocimiento de gestos [21,22]. Las características del dominio tiempo-frecuencia incluyen el modelo autorregresivo (quinto orden), entropía (basada en coeficientes aproximados de wavelet discreta 1D transformada de nivel 4), estimaciones de varianza (basadas en la superposición máxima de ondas discretas transformada de nivel 3), entropía espectral, frecuencia media y potencia de banda. la ondícula el gesto de captar el poder). varias veces en la literatura [10,23,24]. Para facilitar una clasificación efectiva, los extraídos las características se normalizaron a una media de cero restando la media de cada muestra normalización utilizando la desviación estándar.

La extracción de características es crucial para reducir la dimensionalidad de los datos que representan un gesto, permitiendo así la clasificación de los gestos. Un total de 22 características en el dominio del tiempo y la frecuencia de las 8000 muestras de señales EMG segmentadas largas para cada gesto. Las características incluyeron medidas estadísticas como raíz cuadrática media (RMS), varianza (VAR), valor absoluto medio (MAV), cambio de signo de pendiente (SSC), cruces por cero (ZC), (ZC), longitud de la forma de onda (WL), desviación absoluta mediana (MAD), asimetría, curtosis y energía . Estas características están bien establecidas en la literatura por su utilidad en el reconocimiento de gestos [21,22]. Las características del dominio tiempo-frecuencia incluyen el modelo autorregresivo (quinto orden), entropía (basada en coeficientes aproximados de wavelet discreta 1D transformada de nivel 4), estimaciones de varianza (basadas en la superposición máxima de ondas discretas transformada de nivel 3), entropía espectral, frecuencia media y potencia de banda. la ondícula el gesto de captar el poder). varias veces en la literatura [10,23,24]. Para facilitar una clasificación efectiva, los extraídos las características se normalizaron a una media de cero restando la media de cada muestra normalización utilizando la desviación estándar.

La extracción de características es crucial para reducir la dimensionalidad de los datos que representan un gesto, permitiendo así la clasificación de los gestos. Un total de 22 pasos de tiempo y frecuencia para reducir aún más la dimensionalidad del espacio de características, que se derivó de la Las características del dominio se extrajeron de las 8000 muestras de señales EMG segmentadas de largo para los datos EMG segmentados obtenidos por los dos sensores. Con un conjunto de características inicial que comprende cada gesto. Las características incluían medidas estadísticas como la raíz cuadrática media (RMS), 44 características distintas (22 × 2 sensores), se utilizó PCA para transformar estas características en un varianza (VAR), valor absoluto medio (MAV), cambio de signo de pendiente (SSC), cruces por cero (ZC), nuevo conjunto de componentes principales. El objetivo principal era retener la información más informativa. longitud de la forma de onda (WL), desviación absoluta mediana (MAD), asimetría, curtosis y componentes al tiempo que se reduce la dimensionalidad. Las características principales capturan más del 95% de variabilidad ergía. Estas características están bien establecidas en la literatura por su utilidad en el reconocimiento de gestos [21,22]. Las características del dominio tiempo-frecuencia incluyen el modelo autorregresivo.

en todo el conjunto de datos fueron seleccionados meticulosamente. Este criterio de selección aseguró la retención de solo aquellos componentes que contribuyeron significativamente a la variación de los datos, comprimiendo así el espacio de características y preservando su información esencial. Estos componentes seleccionados, que capturaron la mayor parte de la variabilidad del conjunto de datos, se introdujeron exclusivamente en el clasificador. Este uso estratégico de funciones reducidas, aunque informativas, derivadas de PCA tenía como objetivo mejorar el rendimiento de la clasificación centrándose en los aspectos más críticos de los datos de EMG.

2.4. Clasificador de aprendizaje automático

En este artículo, se empleó SVM para clasificar cinco gestos con las manos. Como SVM es un método basado en el kernel, seleccionar las funciones del kernel adecuadas y los hiperparámetros asociados es una tarea importante. Este problema suele resolverse mediante un enfoque de prueba y error. Además, una aplicación SVM típica de un solo núcleo adopta con frecuencia los mismos hiperparámetros para cada clase, y puede no ser adecuada cuando las distribuciones de patrones de características son significativamente diferentes entre las diferentes clases. Aunque existen diferentes núcleos, como el núcleo gaussiano, el núcleo polinómico y el núcleo sigmoide, a menudo no está claro cuál es el núcleo más adecuado para un conjunto de datos determinado y, por lo tanto, es deseable que los métodos del núcleo utilicen una función del núcleo optimizada que Se adapta bien al conjunto de datos disponible y al tipo de límites entre clases. Una forma eficiente de diseñar un núcleo que sea óptimo para un conjunto de datos dado es considerar el núcleo como una combinación convexa de núcleos básicos como se ilustra en la Ecuación (1).

Un SVM basado en MKL está inspirado en [25].

$$K(x, y) = a \cdot K1(x, y) + b \cdot K2(x, y) + c \cdot K3(x, y) + d \cdot KRBF(x, y) \quad (1)$$

Aquí, K1 es el núcleo lineal, K2 es el núcleo cuadrático, K3 es el núcleo cúbico y KRBF es el núcleo de función de base radial o gaussiana con desviación estándar unitaria.
Los coeficientes {a, b, c, d} son hiperparámetros que se ajustan mediante una validación cruzada de búsqueda de cuadrícula de 10 veces. El parámetro de elasticidad o restricción de caja C también es un hiperparámetro que expresa el grado de pérdida de restricción. Una C grande puede clasificar las muestras de entrenamiento de manera más correcta, pero también termina sobreajustando y reduciendo la precisión de las pruebas, por lo tanto, la restricción del cuadro también se ajusta mediante una validación cruzada de búsqueda de cuadrícula de 10 veces. El entrenamiento y las pruebas de los datos se realizan utilizando una división típica de 80 a 20 pruebas de tren. Se utilizó PCA para la selección de características para reducir aún más la dimensionalidad del vector de entrada proporcionado al clasificador y elegir solo características estadísticamente significativas.

Un análisis adicional abarcó la utilización de tres clasificadores de aprendizaje automático: Bayes ingenuo, árbol de decisión y KNN. Esto se llevó a cabo para evaluar el rendimiento de SVM dentro del conjunto de datos específico y las condiciones experimentales, centrándose en la solidez con un número cada vez mayor de gestos y diferentes configuraciones de sensores. Aunque está bien establecido que las SVM son efectivas para el reconocimiento de gestos basado en EMG [26,27], la investigación tiene como objetivo investigar estos hallazgos dentro de nuestra configuración única que utiliza sensores sEMG de bajo costo disponibles comercialmente . Al confirmar la solidez y eficacia de las SVM en esta aplicación, se agregará más evidencia a la teoría establecida, considerando cualquier matiz de nuestro conjunto de datos específico.

3. Resultados

3.1. Evaluación de la configuración del sensor

El mejor conjunto de hiperparámetros que da como resultado la mayor precisión en los datos de prueba para los clasificadores SVM, Bayes ingenuo, KNN y árbol de decisión, creados utilizando un conjunto de datos fusionado de todos los participantes en diferentes configuraciones de sensores, se muestra en la Tabla 1, Tabla 2, Tabla 3 y Tabla 4, respectivamente. Los hiperparámetros asociados se ajustan mediante una validación cruzada de búsqueda de cuadrícula de 10 veces, y las etiquetas utilizadas para cada gesto se muestran en la Figura 1.

Tabla 1. Configuración óptima de hiperparámetros para la clasificación de gestos basada en SVM utilizando 10 veces validación cruzada. El orden del núcleo SVM es el orden del núcleo polinomial puro. {a, b, c, d} son los MKL coeficientes. C es la restricción de caja. C1 , C2 , C3 , C4 , C5 son las cinco configuraciones de sensores diferentes.

	Hiperparámetros		Configuraciones de sensores				
			C1	C2	C3	C4	C5
Clasificando cinco gestos incluyendo pellizco grueso, pellizco fino, agarre de gancho, gesto de señalar y agarre de poder	Orden del núcleo SVM		3	-	-	-	-
	coeficientes MKL	a	-	10	0.1	0.1	1
		b	-	0,5	1	1	0.001
		c	-	0.1	10	10	0
		d	-	0	0	0	0
	C (restricción de caja)		1	0.1	0,5	0,5	1
Clasificando dos gestos incluyendo pellizco fino y agarre potente	Orden del núcleo SVM		1	3	1	2	3
	coeficientes MKL	a	-	-	-	-	-
		b	-	-	-	-	-
		c	-	-	-	-	-
		d	-	-	-	-	-
	C (restricción de caja)		10	1	0,5	0,02	1

Tabla 2. Configuración óptima de hiperparámetros para la clasificación ingenua de gestos basada en Bayes utilizando Validación cruzada 10 veces. El núcleo de distribución 'N' es el núcleo normal y 'B' es el núcleo de caja (uniforme).

	Hiperparámetros	Configuraciones de sensores				
		C1	C2	C3	C4	C5
Clasificando cinco gestos incluyendo pellizco grueso, pellizco fino, agarre de gancho, gesto de señalar y agarre de poder	Núcleo de distribución (NÓTESE BIEN)	norte	B	norte	norte	norte
	Banda ancha	0,05	0,8	0,08	0,15	0,42
Clasificando dos gestos incluyendo pellizco fino y agarre potente	Núcleo de distribución (NÓTESE BIEN)	norte	norte	norte	norte	norte
	Banda ancha	0.1	0,05	0.1	0,05	0,05

Tabla 3. Configuración óptima de hiperparámetros para la clasificación de gestos basada en KNN utilizando 10 veces validación cruzada. D1 a D4 denotan funciones euclidianas, coseno, de manzana y de distancia de Minkowski.

	Hiperparámetros	Configuraciones de sensores				
		C1	C2	C3	C4	C5
Clasificando cinco gestos incluyendo pellizco grueso, pellizco fino, agarre de gancho, gesto de señalar y agarre de poder	Función de distancia	D1	D2	D1	D1	D1
	valor k	1	1	1	1	2
Clasificando dos gestos incluyendo pellizco fino y agarre potente	Función de distancia	D1	D3	D2	D1	D4
	valor k	1	1	1	1	1

Tabla 4. Configuración óptima de hiperparámetros para la clasificación de gestos basada en árbol de decisión utilizando una validación cruzada de 10 veces.

	Hiperparámetros	Configuraciones de sensores				
		C1	C2	C3	C4	C5
Clasificando cinco gestos incluyendo pellizco grueso, pellizco fino, agarre de gancho, gesto de señalar y agarre de poder	Tamaño mínimo de hoja	1	9	2	2	2
Clasificando dos gestos incluyendo pellizco fino y agarre potente	Tamaño mínimo de hoja	2	9	9	2	2

La Tabla 5 presenta los resultados de desempeño de varios clasificadores en dos gestos separados. tareas de clasificación con diferentes configuraciones (C1, C2, C3, C4, C5) y proporciona una descripción general comparativa de la precisión de diferentes clasificadores en cinco distintos y no relacionados. Configuraciones de sensores para tareas de clasificación de gestos. Al clasificar cinco gestos, el SVM El rendimiento del clasificador oscila entre el 75% y el 90%, observándose la mayor precisión en configuración C5. La precisión del KNN fluctúa, alcanzando su punto más alto con 82% con C4 y ligeramente más bajo al 80% con C5. Naïve Bayes muestra su mejor resultado al 70% con C5, mientras que el El clasificador de árbol de decisión tiene un límite del 75% con la misma configuración.

Tabla 5. Exactitud de las pruebas para los clasificadores SVM, KNN, Naïve Bayes y árbol de decisión en todo diferentes configuraciones de sensores. C1 a C5 son las cinco configuraciones de sensores diferentes.

	Clasificador	Configuración Ci				
		C1	C2	C3	C4	C5
Clasificando cinco gestos incluyendo pellizco grueso, pellizco fino, agarre de gancho, gesto de señalar y agarre de poder	SVM	75%	75%	84,6%	87,2%	90%
	knn	67,9%	73,9%	69,2%	82%	80%
	Bayes ingenuo	53,6%	51,3%	61,5%	61,5%	70%
	Árbol de decisión	57,2%	59%	56,4%	59%	75%
Clasificando dos gestos incluyendo pellizco fino y agarre potente	SVM	100%	100%	100%	100%	100%
	knn	100%	86,7%	100%	93,3%	100%
	Bayes ingenuo	100%	73,3%	100%	93,3%	93,3%
	Árbol de decisión	100%	86,7%	100%	80%	100%

Por el contrario, la tarea de clasificar dos gestos, pellizco fino y agarre de fuerza, ve un rendimiento notablemente superior y perfecto por parte del clasificador SVM, manteniendo una precisión del 100%. en todas las configuraciones. Los clasificadores KNN y de árbol de decisión también funcionan excepcionalmente bueno, con KNN logrando una precisión del 100% en todos menos en C2 y C4, donde obtiene una puntuación ligeramente inferior al 86,7% y 93,3%, respectivamente. Los clasificadores Naïve Bayes y árboles de decisión exhibe una precisión perfecta del 100% para la configuración C3. Esta clara diferencia en el desempeño entre tareas sugiere que ciertos clasificadores, especialmente SVM, pueden ser más robustos para cambios en las configuraciones o son más adecuados para tareas de clasificación binaria en el contexto de reconocimiento de gestos.

3.2. Evaluación del clasificador

La matriz de confusión para el clasificador SVM en diferentes configuraciones ha sido Trazado meticulosamente para proporcionar una visualización completa del rendimiento del modelo. (Figura 5).

La matriz de confusión para el clasificador SVM en diferentes configuraciones ha sido Trazado meticulosamente para proporcionar una visualización completa del rendimiento del modelo (Figura 5).

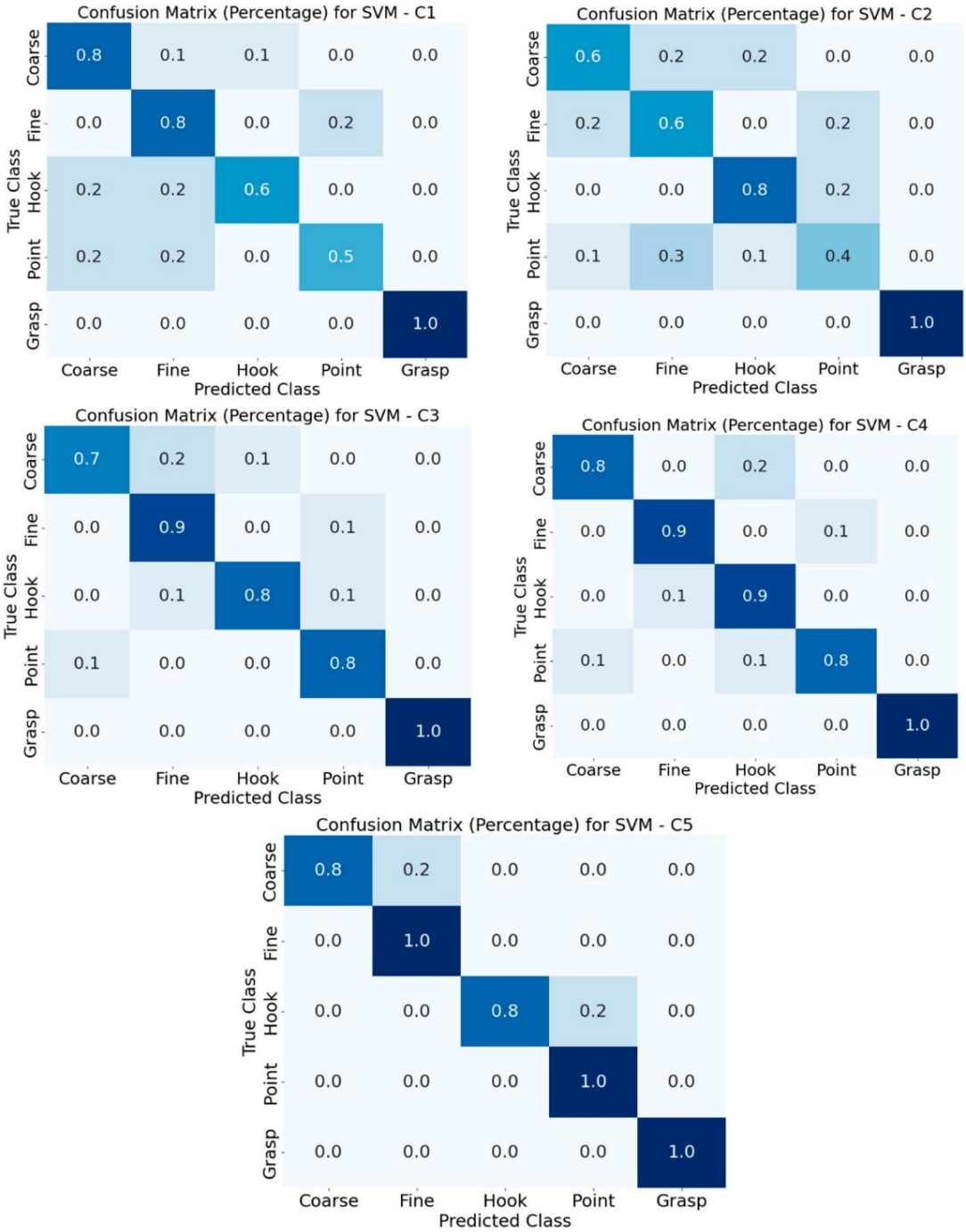


Figura 5. Matriz de confusión para el clasificador SVM en diferentes configuraciones.

Figura 5. Matriz de confusión para el clasificador SVM en diferentes configuraciones.

4. Discusión

El estudio va más allá de la evaluación de la precisión de la clasificación, profundizando en las complejidades de clasificar una gama cada vez mayor de gestos, como lo corroboran investigaciones anteriores [28]. Esta investigación revela además las limitaciones del uso de un solo sensor para la adquisición de datos, que restringe el número de clases claramente identificables. Evaluamos la eficacia de cuatro clasificadores de aprendizaje automático (SVM, Naïve Bayes, KNN y decitree) en cinco configuraciones distintas de sensores para recopilar EMG detallada de gestos con las manos de los antebrazos de ocho participantes. Un amplio conjunto de 22 funciones, que abarcan Tanto el dominio de tiempo como el de frecuencia por sensor se extrajeron de las señales EMG. Previo Los estudios sugieren que un conjunto de características de dominio mixto puede reforzar el rendimiento del clasificador [29].

complejidades de clasificar una gama cada vez mayor de gestos, como lo corroboran investigaciones anteriores [28]. Esta investigación revela además las limitaciones del uso de un solo sensor para la adquisición de datos, que restringe el número de clases claramente identificables. Evaluamos la eficacia de cuatro clasificadores de aprendizaje automático (SVM, Naïve Bayes, KNN y decitree) en cinco configuraciones distintas de sensores para recopilar EMG detallada de gestos con las manos de los antebrazos de ocho participantes. Un amplio conjunto de 22 funciones, que abarcan Tanto el dominio de tiempo como el de frecuencia por sensor se extrajeron de las señales EMG. Previo

Árbol de visión: más de cinco configuraciones de sensores distintas para recopilar datos detallados de

La aplicación de PCA para seleccionar características clave redujo significativamente la dimensionalidad del espacio de características, mejorando la precisión de la clasificación. Los resultados, como se resumen en las Tablas 1 a 5, identificaron la configuración del sensor C5 (dos sensores colocados proximales a la muñeca a lo largo del flexor superficial de los dedos y el flexor largo del pulgar) como el más efectivo, produciendo la mayor precisión con SVM, Bayes ingenuo y árbol de decisión. clasificadores. El clasificador SVM, potenciado por MKL, logró consistentemente más del 90% de precisión en la clasificación de gestos a través de C5, como se muestra en la Tabla 2 y la Figura 5. La configuración de sensor dual de C5, que captura datos de los músculos que controlan los movimientos del pulgar y el índice, sugiere una estrategia ventajosa para la colocación de sensores en el diseño de prótesis de mano.

El gesto de pellizco fino, representado en la Figura 1b, resultó particularmente desafiante debido a sus niveles más bajos de contracción muscular y a la dependencia de solo dos dedos, lo que lo hace más propenso al ruido. El posicionamiento preciso del sensor del C5 sobre los músculos activos durante este gesto permitió la captura de datos más matizados. Una comparación de C1 y C5 (Figura 2) ilustra la influencia significativa de un sensor EMG adicional en la precisión de la clasificación, lo que confirma la superioridad de la EMG de doble canal sobre la de un solo canal en el reconocimiento de gestos.

Cuando las limitaciones de espacio limitan la colocación del sensor cerca de la muñeca en un brazo protésico, C4 (dos sensores colocados proximales a la muñeca a lo largo del flexor superficial de los dedos y a lo largo del flexor superficial de los dedos y el flexor profundo de los dedos) surge como una alternativa factible, logrando una precisión del 87,2% con SVM (Tabla 5). Estos resultados resaltan la importancia tanto de la ubicación del sensor en el músculo como de la proximidad al vientre muscular para una recopilación óptima de datos.

La mejora del rendimiento de SVM por parte de MKL es evidente en la configuración del sensor C3 (un sensor colocado a lo largo del flexor superficial de los dedos y el flexor profundo de los dedos), donde un solo sensor compite estrechamente con el sensor dual C4 en precisión (precisión de la prueba de SVM, Tabla 5). La selección adaptativa de funciones del núcleo de MKL para los datos de C3 es un avance significativo con respecto a los métodos de un solo núcleo. El hiperparámetro C de SVM juega un papel fundamental en el equilibrio de la complejidad del modelo frente al sobreajuste, con un rango de valores de 0,01 a 10, evaluados a través de una validación cruzada de búsqueda de cuadrícula de 10 veces (Tabla 1). Este ajuste fue crucial para desarrollar un modelo SVM optimizado con fuertes capacidades de generalización para una clasificación precisa de gestos. Por lo tanto, para brazos protésicos que solo pueden incorporar un sensor EMG, la configuración recomendada es C3 . El uso de MKL con SVM mejora significativamente el rendimiento sobre las SVM base, especialmente para clasificar múltiples gestos y nos permitió lograr coeficientes MKL distintos de cero para clasificar los cinco gestos como se detalla en la Tabla 1. Mientras que las SVM base sobresalen en la clasificación binaria, MKL maneja la La complejidad de las señales EMG es mejor combinando múltiples núcleos. Esto da como resultado una clasificación sólida de cinco gestos distintos, lo que justifica la complejidad adicional de MKL.

Las matrices de confusión demuestran el alto rendimiento de SVM en diferentes gestos, alrededor del 80% al 100% con la configuración C5 . Además, SVM se destacó como clasificador binario para los gestos de pellizco fino versus agarre fuerte (Figura 1), logrando una precisión del 100 %. Una tendencia observada es la reducción de la precisión de la clasificación con un número creciente de gestos , un fenómeno que resuena con hallazgos anteriores [26]. Esta disminución se atribuye a la dispersión más amplia de las señales EMG en el antebrazo y la superposición de señales resultante de las contracciones musculares simultáneas al realizar gestos complejos.

Una posible limitación del presente estudio, cuando se extiende a la clasificación EMG en línea en tiempo real, es la ligera variación en las longitudes de los segmentos EMG en el dominio temporal utilizados para la extracción de características en diferentes configuraciones de gestos. Esta limitación se puede abordar fácilmente seleccionando longitudes de segmentos exactamente iguales durante la solicitud en línea. Otra limitación del trabajo actual es el tamaño de la muestra de ocho participantes. Este tamaño de muestra es similar a otros tamaños de muestra en la literatura [25,27]. Sin embargo, el tamaño limitado de la muestra puede afectar la solidez de las precisiones obtenidas hasta cierto nivel. El estudio actual sirve para validar la metodología propuesta a pequeña escala, y está previsto realizar experimentos futuros con un tamaño de muestra más grande e incluir individuos con diferencias en las extremidades para evaluar directamente la aplicación de nuestros hallazgos en el control de prótesis de mano. Esto mejorará la generalización y relevancia de la investigación actual.

Investigaciones futuras aplicarán los hallazgos de este artículo para replicar los gestos de las manos en una prótesis de mano mioeléctrica en tiempo real. También explorará diferentes configuraciones de sensores y técnicas de aprendizaje automático para mejorar aún más el rendimiento de las prótesis mioeléctricas. Además, el desarrollo de métodos más sofisticados de extracción de características y optimización del clasificador podría ser beneficioso para manejar el análisis de señales EMG multifacético. Este estudio sirve como un trampolín hacia la realización de prótesis mioeléctricas más eficientes y efectivas, y se espera que los conocimientos adquiridos inspiren una mayor exploración en este campo prometedor.

5. Conclusiones

Este estudio ha logrado avances significativos en el campo de la tecnología de prótesis mioeléctricas, con hallazgos clave que tienen el potencial de dar forma a futuras investigaciones y aplicaciones. El rendimiento superior de la configuración de anillo índice y pulgar índice (C4 y C5) de configuración de doble canal subraya la importancia de una ubicación óptima del sensor para mejorar la funcionalidad de las prótesis mioeléctricas. La aplicación de MKL en el clasificador SVM, particularmente al colocar un sensor en anillo (configuración C3), ha demostrado su eficacia para lograr una alta precisión de clasificación incluso con un solo sensor. La disminución observada en la precisión con un aumento en el número de gestos resalta la necesidad de una extracción integral de características y una optimización del clasificador para el análisis complejo de señales EMG. Estos hallazgos en conjunto allanan el camino para avances en prótesis mioeléctricas, enfatizando la necesidad de configuraciones de sensores refinadas y metodologías de aprendizaje automático para mejorar la precisión del reconocimiento de gestos.

Contribuciones de los autores: Conceptualización, LC, SH, KBM, JA y HR; Metodología, LC, SH, KBM, JA y HR; Software, LC y SH; Validación, LC, SH, KBM y HR; Análisis formal , LC, SH, JA y HR; Investigación, LC, SH, KBM, JA y HR; Recursos, KBM, JA y RRHH; Curación de datos, LC, SH, KBM y HR; Redacción: borrador original, LC, SH, KBM, JA y HR; Redacción: revisión y edición, LC, SH, KBM, JA y HR Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: Esta investigación fue financiada por la Glenrose Hospital Foundation y Mitacs, número de subvención: IT36690.

Declaración de la Junta de Revisión Institucional: El estudio se realizó de acuerdo con la Declaración de Helsinki y fue aprobado por el Comité de Ética Institucional de la Universidad de Alberta (AB T6G 2N2, aprobado el 25 de abril de 2022).

Declaración de consentimiento informado: Todos los participantes firmaron un formulario de consentimiento, el procedimiento fue aprobado por el número de solicitud de la junta de ética de investigación de la Universidad de Alberta: Pro00117863.

Declaración de disponibilidad de datos: Datos disponibles previa solicitud.

Conflictos de intereses: Los autores declaran no tener conflictos de intereses.

Referencias

1. McDonald, CL; Westcott-McCoy, S.; Tejedor, señor; Haagsma, J.; Kartin, D. Prevalencia global de miembros traumáticos no fatales amputación. Prótesis. Ortot. En t. 2021, 45, 105-114. [Referencia cruzada] [PubMed]
2. Efánov, J.; Tchiloemba, B.; Izadpanah, A.; Harris, P.; Danino, M. Una revisión de las utilidades y costos del tratamiento de amputaciones de extremidades superiores con alotrasplante compuesto vascularizado versus prótesis mioeléctricas en Canadá. Abierto JPRAS 2022, 32, 150–160. [Referencia cruzada] [PubMed]
3. Guía del usuario de Hero Arm: Open Bionics. Disponible en línea: https://openbionics.com/hero-arm-user-guide/ (consultado el 9 diciembre de 2023).
4. bebionic|Ottobock Estados Unidos. Disponible en línea: https://www.ottobockus.com/prosthetics/upper-limb-prosthetics/solution-overview/mano-bebiónica/ (consultado el 9 de diciembre de 2023).
5. Peerdeman, B.; Boere, D.; Witteveen, H.; Hermens, H.; Stramigioli, S.; Rietman, H.; Veltink, P.; Misra, S. Prótesis mioeléctricas de antebrazo: estado del arte desde una perspectiva centrada en el usuario. J. Rehabilitación. Res. Desarrollo. 2011, 48, 719–738. [Referencia cruzada] [PubMed]
6. Cordella, F.; Ciancio, AL; Sacchetti, R.; Davalli, A.; Cutti, AG; Guglielmelli, E.; Zollo, L. Revisión de la literatura sobre las necesidades de la parte superior. Usuarios de prótesis de extremidades. Frente. Neurociencias. 2016, 10, 209. [Referencia cruzada] [PubMed]
7. Geethanjali, P. Control mioeléctrico de prótesis de manos: revisión del estado del arte. Medicina. Dispositivos 2016, 9, 247–255. [Referencia cruzada] [PubMed]

8. Tavakoli, M.; Benussi, C.; Lopes, PA; Osorio, LB; de Almeida, AT Reconocimiento robusto de gestos manuales con un brazalete portátil EMG de superficie de doble canal y clasificador SVM. Biomédica. Proceso de señal. Control. 2018, 46, 121-130. [Referencia cruzada]
9. Palkowski, A.; Redlarski, G. Clasificación básica de gestos manuales basada en electromiografía de superficie. Computadora. Matemáticas. Métodos Med. 2016, 2016, 6481282. [Referencia cruzada] [PubMed]
10. Lee, KH; Min, JY; Byun, S. Clasificación basada en electromiograma de gestos con manos y dedos utilizando redes neuronales artificiales. Sensores 2021, 22, 225. [CrossRef] [PubMed]
11. MAhsan, R.; Ibrahimy, MI; Khalifa, OO Reconocimiento de gestos con las manos basado en señales de electromigrafía (EMG) utilizando una red neuronal artificial (ANN). En Actas de la Cuarta Conferencia Internacional sobre Mecatrónica de 2011: Ingeniería integrada para el desarrollo industrial y social, ICOM'11 — Actas de la conferencia, Kuala Lumpur, Malasia, 17 a 19 de mayo de 2011. [CrossRef]
12. Kisa, DH; Ozdemir, MA; Guren, O.; Akan, A. Clasificación de gestos con las manos basada en EMG mediante series temporales de descomposición en modo empírico y aprendizaje profundo. En Actas del TIPTEKNO 2020—Tip Teknolojileri Kongresi—2020 Medical Technologies Congress, TIPTEKNO 2020, Antalya, Turquía, 19 y 20 de noviembre de 2020. [CrossRef]
13. Sensor muscular MYOWARE® 2.0. Disponible en línea: https://myoware.com/products/muscle-sensor/ (consultado el 9 de diciembre de 2023). 14. del Toro, SF; Wei, Y.; Olmeda, E.; Ren, L.; Guowu, W.; Díaz, V. Validación de un sistema de electromiografía (EMG) de bajo costo mediante un Dispositivo EMG comercial y preciso: estudio piloto. Sensores 2019, 19, 5214. [CrossRef] [PubMed]
15. Heywood, S.; Pua, YH; McClelland, J.; Geigle, P.; Rahmann, A.; Bower, K.; Clark, R. Electromiografía de bajo costo: validación frente a un sistema comercial que utiliza umbrales de tiempo de activación tanto manuales como automatizados. J. Electromiogr. Kinesiol. 2018, 42, 74–80. [Referencia cruzada] [PubMed]
16. Kurniawan, SR; Pamungkas, D. Sensores de brazalete MYO y algoritmo de red neuronal para controlar el robot manual. En Actas de la Conferencia Internacional sobre Ingeniería Aplicada de 2018, ICAE 2018, Batam, Indonesia, 3 y 4 de octubre de 2018. [Referencia cruzada]
17. Mao, Z.-H.; Lee, H.-N.; Sclabassi, RJ; Sun, M. Capacidad informativa del pulgar y el índice en la comunicación. Traducción IEEE . Biomédica. Ing. 2009, 56, 1535-1545. [Referencia cruzada] [PubMed]
18. Hioki, M.; Kawasaki, H. Estimación de los ángulos de las articulaciones de los dedos a partir de sEMG utilizando una red neuronal que incluye el factor de retardo de tiempo y Estructura recurrente. Rehabilitación ISRN. 2012, 2012, 604314. [Referencia cruzada]
19. Experimento: Clasificación de señales. Disponible en línea: https://backyardbrains.com/experiments/RobotHand (consultado el 9 diciembre de 2023).
20. Ghalyan, FIP; Abouelenin, ZM; Annamalai, G.; Kapila, V. Filtro de suavizado gaussiano para mejorar el modelado de señales EMG. En Procesamiento de señales en medicina y biología: tendencias emergentes en investigación y aplicaciones; Springer: Cham, Suiza, 2020; págs. 161-204. [Referencia cruzada]
21. Jaramillo-Yánez, A.; Benalcázar, ME; Mena-Maldonado, E. Reconocimiento de gestos con las manos en tiempo real mediante electromiografía de superficie y aprendizaje automático: una revisión sistemática de la literatura. Sensores 2020, 20, 2467. [CrossRef] [PubMed]
22. Tkach, D.; Huang, H.; Kuiken, TA Estudio de la estabilidad de las características del dominio del tiempo para el reconocimiento de patrones electromiográficos. J. Neuroeng. Rehabilitación. 2010, 7, 21. [Referencia cruzada] [PubMed]
23. Khairuddin, IM; Sidek, SN; Majeed, APPA; Razman, MAM; Puzi, AA; Yusof, HM La clasificación de la intención del movimiento a través de modelos de aprendizaje automático: la identificación de características EMG significativas en el dominio del tiempo. Computación entre pares. Ciencia. 2021, 7, e379. [Referencia cruzada] [PubMed]
24. Tomaszewski, J.; Amaral, T.; Días, O.; Wołczowski, A.; Kurzy´nski, M. Clasificación de señales EMG mediante red neuronal con AR coeficientes del modelo. Procedimiento de la IFAC. vol. 2009, 42, 318–325. [Referencia cruzada]
25. Ella, Q.; Luo, Z.; Meng, M.; Xu, P. Clasificación de patrones EMG basada en SVM con aprendizaje de núcleo múltiple para el control de las extremidades inferiores. En Actas de la 11.ª Conferencia Internacional sobre Control, Automatización, Robótica y Visión, ICARCV 2010, Singapur, 7 a 10 de diciembre de 2010; págs. 2109-2113. [Referencia cruzada]
26. Odeyemi, J.; Ogbeyemi, A.; Wang, K.; Zhang, W. Sobre el agarre automatizado de objetos para prótesis de manos inteligentes mediante el aprendizaje automático. Bioingeniería 2024, 11, 108. [CrossRef] [PubMed]
27. Avilés, M.; Sánchez-Reyes, L.-M.; Fuentes-Aguilar, RQ; Toledo-Pérez, DC; Rodríguez-Reséndiz, J. Una novedosa metodología para clasificar movimientos EMG basada en SVM y algoritmos genéticos. Micromáquinas 2022, 13, 2108. [CrossRef] [PubMed]
28. Fajardo, JM; Gómez, O.; Prieto, F. Clasificación EMG de los gestos de las manos utilizando rasgos profundos y artesanales. Biomédica. Proceso de señal. Control. 2021, 63, 102210. [Referencia cruzada]
29. Abbaspour, S.; Lindén, M.; Gholamhosseini, H.; Naber, A.; Ortiz-Catalan, M. Evaluación de algoritmos de reconocimiento basados en EMG de superficie para decodificar movimientos de la mano. Medicina. Biol. Ing. Computadora. 2020, 58, 83–100. [Referencia cruzada] [PubMed]

electrónica

MDPI

Artículo

Mejora de la utilidad de datos en datos de ubicación que preservan la privacidad Colección a través de partición de cuadrícula adaptable

Jongwook Kim

Departamento de Ciencias de la Computación, Universidad Sangmyung, Seúl 03016, República de Corea; jkim@smu.ac.kr

Resumen: La disponibilidad generalizada de dispositivos habilitados para GPS y los avances en las tecnologías de posicionamiento han facilitado significativamente la recopilación de datos de ubicación de los usuarios, convirtiéndolos en un activo invaluable en diversas industrias. Como resultado, existe una demanda creciente de recopilación e intercambio de estos datos. Dada la naturaleza sensible de la información de ubicación del usuario, se han realizado esfuerzos considerables para garantizar la privacidad, y los esquemas basados en privacidad diferencial (DP) emergen como el enfoque más preferido . Sin embargo, estos métodos normalmente representan ubicaciones de usuarios en cuadrículas divididas uniformemente, que a menudo no reflejan con precisión la verdadera distribución de los usuarios dentro de un espacio. Por lo tanto, en este artículo, presentamos un método novedoso que ajusta de forma adaptativa la cuadrícula en tiempo real durante la recopilación de datos, representando así a los usuarios en estas cuadrículas divididas dinámicamente para mejorar la utilidad de los datos recopilados. Específicamente, nuestro método captura directamente la distribución de usuarios durante el proceso de recopilación de datos, eliminando la necesidad de depender de datos de distribución de usuarios preexistentes. Los resultados experimentales con conjuntos de datos reales muestran que el esquema propuesto mejora significativamente la utilidad de los datos de ubicación recopilados en comparación con el método existente.

Palabras clave: privacidad de ubicación; distribución de densidad; privacidad diferencial; geo-indistinguibilidad

Cita: Kim, J. Mejora de los datos Utilidad en la recopilación de datos de ubicación que preserva la privacidad mediante partición de red adaptable. Electrónica 2024, 13, 3073. https://doi.org/10.3390/electrónica13153073

Editores académicos: Swapnoneel Roy y Guosheng Xu

Recibido: 27 de junio de 2024
Revisado: 27 de julio de 2024
Aceptado: 29 de julio de 2024
Publicado: 3 de agosto de 2024

1. Introducción

La proliferación de dispositivos con GPS y los recientes avances en las tecnologías de posicionamiento han facilitado la recopilación de datos de ubicación de los usuarios, convirtiéndolos en un activo valioso para diversos sectores. Estos datos desempeñan un papel importante en áreas como el marketing personalizado, análisis de tráfico en tiempo real, recomendaciones, etc. Por ejemplo, el análisis de tráfico en tiempo real utiliza datos de ubicación para optimizar el flujo de tráfico, reducir la congestión y mejorar los sistemas de navegación para viajes más eficientes. [1,2]. Además, las recomendaciones basadas en la ubicación para servicios, restaurantes y eventos brindan a los usuarios sugerencias relevantes y oportunas, mejorando su experiencia general [3,4]. Como resultado, la demanda de recopilar y compartir datos de ubicación de los usuarios sigue aumentando.

Los datos de ubicación del usuario son confidenciales porque contienen información personal, como direcciones de casa o de la empresa, registros de visitas al hospital e incluso afiliaciones políticas [5–7]. Por ejemplo, al recopilar y analizar la información de posicionamiento de los visitantes en un gran centro comercial cubierto, es posible inferir detalles sensibles, como sus patrones de compra. Además, los datos de ubicación se pueden comparar con otros conjuntos de datos para sacar conclusiones aún más precisas sobre el estilo de vida y las elecciones de un individuo [8]. Por ejemplo, las visitas frecuentes a ciertos tipos de negocios o lugares pueden indicar condiciones de salud, pasatiempos o incluso prácticas religiosas específicas. Como resultado, la recopilación indiscriminada de datos de ubicación plantea importantes preocupaciones sobre la privacidad. En consecuencia, se han realizado esfuerzos considerables para proteger la privacidad de los datos de ubicación de los usuarios al manejar dichos datos.

A medida que la privacidad diferencial (DP) [9,10] se ha convertido en el estándar de facto para el manejo de datos personales sensibles, se han realizado importantes esfuerzos para aplicarla a los datos de ubicación. Como resultado, se han propuesto numerosos métodos basados en DP para recopilar, procesar y analizar datos de ubicación preservando al mismo tiempo la privacidad. Muchos de estos enfoques representan datos de ubicación del usuario mediante cuadrículas, donde todo el dominio se divide uniformemente en elementos disjuntos.

cuadrículas, y la ubicación de un usuario está representada por la cuadrícula en la que se encuentra su posición real [11-14]. Aunque representar la ubicación del usuario mediante cuadrículas divididas uniformemente es sencillo, no tiene en cuenta la distribución real de los usuarios dentro del espacio.
Este enfoque a menudo resulta en una menor utilidad de los datos de ubicación recopilados, ya que supone que los usuarios están distribuidos uniformemente en toda el área. Sin embargo, esto no es cierto en la mayoría de los escenarios del mundo real, donde algunas áreas son más densas que otras. Por ejemplo, en un entorno urbano, el centro de la ciudad puede tener una alta concentración de usuarios, mientras que los suburbios tienen una menor densidad. Esta discrepancia puede afectar negativamente la precisión y eficacia de los análisis posteriores. Por lo tanto, se necesitan representaciones de cuadrícula más sofisticadas que se alineen con las distribuciones reales de los usuarios para mejorar la utilización de los datos y mejorar el rendimiento de estas aplicaciones.

Las soluciones existentes suponen la existencia de información previa sobre la distribución de usuarios, normalmente obtenida a partir de datos históricos. Sin embargo, es posible que dichos datos históricos no siempre estén disponibles para muchas aplicaciones. Más importante aún, la información previa sobre la distribución de usuarios derivada de datos históricos puede no coincidir con la distribución actual, ya que puede cambiar con el tiempo o en respuesta a eventos sociales especiales. Por ejemplo, eventos importantes como festivales pueden cambiar drásticamente los patrones y densidades de movimiento de los usuarios, haciendo que los datos históricos queden obsoletos o sean engañosos [15]. Por lo tanto, es preferible extraer instantáneamente información sobre la distribución de usuarios durante la recopilación de datos y ajustar de forma adaptativa la cuadrícula en consecuencia.

En este artículo, proponemos un método novedoso que extrae simultáneamente la distribución de usuarios y ajusta de forma adaptativa la cuadrícula en tiempo real durante la recopilación de datos de ubicación. Las aportaciones de este trabajo se pueden resumir en:

- Primero, presentamos un método para calcular de manera efectiva la distribución de usuarios durante la recopilación de datos de ubicación basada en DP. Este enfoque es capaz de capturar eficazmente la distribución de los usuarios en tiempo real y adaptarse a los cambios dinámicos en el comportamiento de los usuarios.
- Luego, proponemos un método para ajustar de forma adaptativa la cuadrícula para maximizar la utilidad de los datos de ubicación recopilados bajo DP. Este ajuste de cuadrícula adaptativo está diseñado para mejorar la granularidad y relevancia de los datos, asegurando que se prioricen las áreas más importantes y densamente pobladas, mejorando así la calidad general y la aplicabilidad de los datos.
- Evaluamos el rendimiento de los algoritmos propuestos utilizando conjuntos de datos del mundo real. Los resultados de la evaluación demostraron que el esquema propuesto mejora significativamente la utilidad de los datos de ubicación recopilados en comparación con los métodos existentes.

El resto de este documento está organizado de la siguiente manera: La Sección 2 revisa el trabajo relacionado. En la Sección 3, proporcionamos información general. En la Sección 4, presentamos un método novedoso que extrae simultáneamente la distribución de los usuarios y ajusta de forma adaptativa la cuadrícula en tiempo real durante la recopilación de datos de ubicación. En la Sección 5, evaluamos experimentalmente el enfoque propuesto con conjuntos de datos reales. Finalmente, la Sección 6 presenta nuestras conclus

2. Trabajo relacionado

Se han desarrollado numerosos métodos basados en DP para recopilar, procesar y analizar datos de ubicación preservando al mismo tiempo la privacidad. En esta sección, proporcionamos una breve descripción general de estos métodos.

La privacidad diferencial local (LDP) es una variante de DP en la que cada usuario altera individualmente sus propios datos confidenciales antes de informarlos al servidor. Kim y Jang [12] proponen un enfoque de agregación de datos basado en LDP diseñado para la recopilación de datos de posicionamiento en interiores teniendo en cuenta la carga de trabajo, garantizando al mismo tiempo la privacidad del usuario. Su método identifica una estrategia óptima de codificación y perturbación de datos dentro del marco LDP para minimizar el error de estimación general para la carga de trabajo dada. LDPTrace [14] está diseñado para generar sintéticamente datos de trayectoria locales diferencialmente privados. En este método, la información de ubicación del usuario se recopila mediante LDP para garantizar la privacidad, y estos datos perturbados luego se utilizan para generar trayectorias sintéticas. Kim y cols. [3] presentan un método para recomendar el siguiente punto de interés, utilizando datos de ubicación recopilados bajo LDP.

La privacidad diferencial métrica (MDP) amplía el marco de privacidad diferencial estándar para manejar datos con medidas métricas o de distancia inherentes [16]. Esta extensión es particularmente útil para datos basados en la ubicación. La geoindistinguibilidad (Geo-Ind) es una aplicación específica de MDP diseñada para servicios basados en ubicación [11,17,18]. Los marcos de Mobile Crowdsens-ing (MCS) a menudo utilizan Geo-Ind para recopilar información de ubicación de los trabajadores y asignar tareas de manera que se preserve la privacidad. Wang y cols. [19] fue el primero en utilizar Geo-Ind para proteger la privacidad de la ubicación de los trabajadores en el proceso MCS. Su marco propuesto incluye tres pasos: Primero, el servidor MCS genera una función que satisface Geo-Ind. A continuación, cada trabajador descarga esta función, ofusca su verdadera ubicación y carga la ubicación ofuscada en el servidor. Finalmente, el servidor MCS asigna tareas a los trabajadores basándose en la información de ubicación ofuscada. En [20], se investiga la protección de la privacidad de la ubicación en MCS basados en vehículos, donde la hoja de ruta se modela como un gráfico dirigido ponderado con ubicaciones de tareas y trabajadores como puntos en el gráfico. Los autores proponen un esquema de ofuscación basado en un mecanismo de optimización que logra la ofuscación de la ubicación a través de una distribución probabilística sobre el gráfico que satisface Geo-Ind. Jin et al. [21] propone un marco de comercio de privacidad de ubicación centrado en el usuario para MCS. Siguiendo la noción de Geo-Ind, diseñan un mecanismo de ofuscación de ubicación que permite a cada trabajador ofuscar probabilísticamente su verdadera ubicación utilizando su propio presupuesto de privacidad. Zhang et al. [13] introduce un método de ofuscación que satisface a Geo-Ind para recopilar información de ubicación de los trabajadores en MCS. Huang et al. [22] proponen un esquema consciente de la privacidad para el monitoreo de ruido basado en MCS, donde el servidor publica tareas y los trabajadores informan ubicaciones perturbadas y niveles de ruido bajo DP. Cada trabajador colabora con un maestro, cuidadosamente seleccionado entre los trabajadores del mismo grupo, para lograr Geo-Ind a nivel de grupo. Zhao y cols. [23] exploraron la protección de la privacidad de las ubicaciones de los individuos en el contexto del análisis de la distribución direccional geográfica de la comunidad. Definieron la información de la comunidad utilizando una matriz de covarianza y la integraron en la indistinguibilidad de la geoelipse propuesta basada en Geo-Ind. Esta indistinguibilidad de geoelipse proporciona garantías de privacidad cuantificables para ubicaciones dentro del espacio de Mahalanobis. Yu et al. [24] resaltaron las debilidades de los actuales mecanismos de ofuscación de ubicación basados en Geo-Ind, especialmente cuando los usuarios comparten consistentemente sus ubicaciones con múltiples proveedores de LBS durante un largo período de tiempo. Para abordar este problema, introdujeron PrivLocAd, un sistema que utiliza perfiles de ubicación para generar ubicaciones ofuscadas, protegiendo así la privacidad del usuario contra ataques adversarios multiplataforma . Zhao y cols. [25] introdujo un nuevo concepto de privacidad llamado indistinguibilidad de vectores, que se basa en Geo-Ind para proporcionar una garantía de privacidad para las relaciones dependientes de la ubicación.

Han desarrollado cuatro mecanismos para lograr la indistinguibilidad de los vectores, utilizando distribuciones uniformes y de Laplace. Mendes et al. [26] utilizaron la velocidad del usuario y la frecuencia de informes para medir la correlación entre ubicaciones. Ampliaron Geo-Ind para mejorar la preservación de la privacidad en escenarios de informes continuos en línea. Específicamente, introdujeron un Geo-Ind consciente de la velocidad que equilibra automáticamente la privacidad y la utilidad en función de la velocidad del usuario y la frecuencia de los informes de ubicación.

EGeoIndis [27] es un marco de protección de la privacidad de la ubicación de vehículos diseñado para la estimación de la densidad del tráfico. Aprovecha Geo-Ind para proteger la privacidad de la ubicación del vehículo durante el proceso de estimación de la densidad del tráfico. En [28], los autores propusieron un método basado en aprendizaje profundo para estimar la distribución de densidad utilizando datos de ubicación recopilados bajo Geo-Ind. Chen et al. [29] desarrollan un método para crear un mapa de vulnerabilidad de COVID-19 utilizando la distribución de densidad de participantes voluntarios con síntomas de COVID-19. Explotan Geo-Ind para recopilar las ubicaciones de los participantes de una manera que preserve la privacidad y garantice la confidencialidad de la información de salud confidencial. Fathalizadeh et al. [30] presentan un marco para implementar Geo-Ind para ambientes interiores. El marco propuesto considera dos escenarios para aplicar Geo-Ind, informando un punto ofuscado al proveedor de servicios de ubicación que satisface DP.

El enfoque propuesto en este artículo aprovecha Geo-Ind, que es un modelo representativo en el ámbito de la recopilación de datos de ubicación que preservan la privacidad. Sin embargo, el enfoque propuesto difiere de otros métodos basados en Geo-Ind en varios aspectos. primero, existente

Los métodos suelen depender de la disponibilidad de datos históricos para inferir la distribución de los usuarios, lo que presenta desafíos importantes. Es posible que los datos históricos no siempre sean accesibles o no estén actualizados, lo que da lugar a inferencias inexactas sobre la distribución de los usuarios. En segundo lugar, la mayoría de los métodos existentes utilizan estructuras de cuadrícula estáticas, lo que da como resultado una representación fija que no tiene en cuenta el movimiento dinámico de los usuarios. Por el contrario, el método propuesto aborda estas limitaciones ajustando de forma adaptativa las cuadrículas en tiempo real durante la recopilación de datos. Este ajuste dinámico captura la distribución actual de usuarios sin depender de datos históricos, mejorando así la utilidad de los datos recopilados.

3. Antecedentes y planteamiento del problema

En esta sección, proporcionamos los antecedentes necesarios para este artículo y exponemos el problema abordado en este documento.

3.1. Antecedentes

Recientemente, DP ha surgido como el estándar de facto para el procesamiento de datos que preserva la privacidad . DP se basa en una definición matemática formal que proporciona una garantía de privacidad probabilística contra atacantes con conocimientos previos arbitrarios [9]. Garantiza que un atacante no pueda determinar con gran confianza si un individuo determinado está incluido en los datos difundidos. DP se define formalmente como [9,10]:

Definición 1. (-DP) Un algoritmo aleatorio A satisface -DP, si y solo si para (1) dos conjuntos de datos vecinos cualesquiera, D1 y D2, y (2) cualquier salida O de A, se cumple lo siguiente:

$$\Pr[A(D1) = O] \leq mi \times \Pr[A(D2) = O]. \quad (1)$$

Dos conjuntos de datos, D1 y D2, se consideran vecinos si difieren en un solo registro.
La definición anterior indica que, para cualquier salida de A, un adversario con cualquier conocimiento previo no puede determinar de manera confiable si D1 o D2 fue la entrada a A. El parámetro , conocido como presupuesto de privacidad, regula el nivel de privacidad: más pequeño Los valores proporcionan una protección de la privacidad más fuerte pero agregan más ruido al resultado, mientras que los valores de más grandes ofrecen una protección de la privacidad más débil con menos ruido.

Ha habido varias propuestas para aplicar el concepto de DP a la protección de datos de ubicación. En este artículo, utilizamos Geo-Ind, un concepto basado en el marco de DP bien establecido y reconocido como la definición de privacidad estándar para proteger los datos de ubicación en servicios basados en la ubicación [11,17,18]. Además de los datos de ubicación, Geo-Ind también se utiliza para recopilar otros tipos de datos, como microdatos de texto, de manera compatible con DP [31,32].
Geo-Ind se define formalmente de la siguiente manera:

Definición 2. (-Geo-Ind) Considere X como el conjunto de posibles ubicaciones del usuario e Y como el conjunto de ubicaciones informadas, que normalmente se supone que son iguales. Sea K un mecanismo aleatorio que genera una ubicación perturbada a partir de la ubicación real de un usuario. Un mecanismo aleatorio K satisface -Geo-Ind si y sólo si la siguiente condición se cumple para (1) todos los x1, x2 X y (2) cualquier ubicación de salida y Y:

$$K(x1)(y) \leq e \times K(x2)(y), \text{ donde } d(x1, x2) \quad (2)$$

x2) corresponde a la distancia entre x1 y x2.

Hay dos métodos principales para implementar Geo-Ind: el mecanismo de Laplace y el mecanismo basado en matrices. Es bien sabido que el mecanismo basado en matrices es más eficaz que el método de Laplace, dada la información previa sobre la distribución de usuarios que se puede obtener a partir de los datos históricos disponibles [11]. Esta mayor efectividad se debe a que el mecanismo matricial incorpora información de distribución previa al perturbar las verdaderas ubicaciones de los usuarios. Como resultado, la distribución de ubicaciones perturbadas recopiladas mediante el mecanismo basado en matrices se aproxima más a la distribución real que la distribución recopilada mediante el mecanismo de Laplace.

En el mecanismo basado en matrices, el espacio primero se divide en un conjunto de cuadrículas y luego el servidor de recopilación de datos calcula una matriz de ofuscación, M, sobre estas cuadrículas que satisfacen -Geo-Ind. Esta matriz luego se distribuye a los usuarios. Posteriormente, los usuarios alteran sus datos de ubicación de acuerdo con las probabilidades incorporadas en M e informan la ubicación perturbada al servidor en lugar de sus datos verdaderos. En la literatura se han propuesto varios enfoques para calcular la matriz de ofuscación que satisface -Geo-Ind [11,13,14,32,33]. Sin embargo, observamos que el método propuesto en este artículo es lo suficientemente general como para aplicarse a cualquier mecanismo basado en matrices.

3.2. Planteamiento del problema

Sea U = {u1, u2, · · · , uk} un conjunto de usuarios que aceptan proporcionar información de su ubicación al servidor. Sin embargo, los usuarios no confían plenamente en el servidor y, por lo tanto, en lugar de proporcionar información de ubicación verdadera, cada usuario proporciona información de ubicación perturbada (y por lo tanto preservada la privacidad) que satisface -Geo-Ind. Supongamos que toda el área está dividida en cuadrículas disjuntas y sea G el conjunto de estas cuadrículas. La ubicación de cada usuario se representa entonces mediante la cuadrícula en G a la que pertenece su verdadera ubicación.

El problema abordado en este documento es recopilar datos de ubicación de alta utilidad y al mismo tiempo proteger la privacidad de la ubicación de los usuarios con -Geo-Ind. Los métodos existentes utilizan particiones de red estáticas que no se adaptan a los cambios en tiempo real en la distribución de usuarios o se basan en datos de distribución de usuarios preexistentes, que pueden no estar disponibles o no ser precisos en escenarios en tiempo real. Para abordar estas brechas, proponemos un novedoso método de partición de cuadrícula adaptativa que ajusta dinámicamente la cuadrícula durante el proceso de recopilación de datos de ubicación. En particular, el método propuesto captura directamente la distribución de los usuarios durante la recopilación de datos, eliminando la necesidad de información de distribución preexistente.

4. Método propuesto La

Figura 1 proporciona una descripción general del esquema de recopilación de datos de ubicación propuesto utilizando

-Geo-Ind. • Recopilación de datos de ubicación perturbada de usuarios muestreados: el servidor primero calcula la matriz de ofuscación, M, sobre cuadrículas divididas uniformemente y la distribuye a los usuarios muestreados, quienes luego envían sus ubicaciones perturbadas de regreso al servidor.
• Estimación de la distribución de usuarios: El servidor estima la distribución de usuarios basado en los datos de ubicación perturbada recopilados de los usuarios muestreados. • Cálculo de grids con particiones adaptativas: el servidor utiliza la distribución estimada de usuarios para calcular grids con particiones adaptativas. • Recopilación de datos de ubicación perturbada de los usuarios restantes mediante el uso de cuadrículas con particiones adaptativas: se calcula una nueva matriz de ofuscación utilizando las cuadrículas con particiones adaptativas. Esta nueva matriz de ofuscación se utiliza luego para recopilar datos de ubicación de los usuarios restantes.

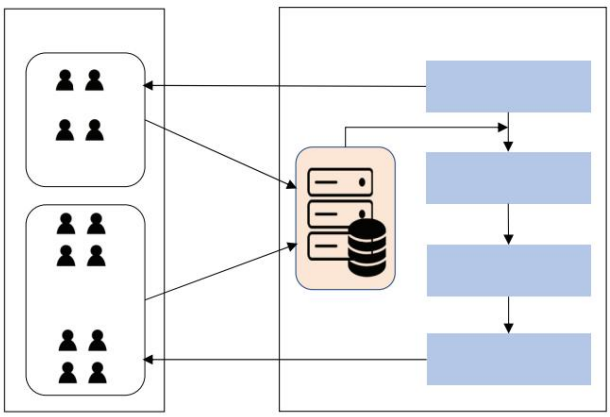


Figura 1. Una descripción general del esquema de recopilación de datos de ubicación propuesto para preservar la privacidad.

En las siguientes subsecciones, proporcionamos una explicación detallada de cada paso.

4.1. Recopilación de datos de ubicación perturbados de usuarios muestreados

Supongamos que toda el área está dividida uniformemente en m cuadrículas, G = {g1, g2, · · · , gm}. El servidor de recopilación de datos calcula una matriz de ofuscación, M, sobre G. Existen varios enfoques para calcular M que satisfacen -Geo-Ind. En este artículo, utilizamos el método propuesto en [13], donde la matriz de ofuscación se define como una matriz m × m . Cada elemento M[i, j], que representa la probabilidad de que una ubicación perturbada gj se genere aleatoriamente a partir de la ubicación verdadera gi , se define de la siguiente manera:

$$M[i, j] = \frac{e^{\frac{\varepsilon}{2}\cdot d(g_i, g_j)}}{\sum_{g_k \in G} e^{\frac{\varepsilon}{2}\cdot d(g_i, g_k)}} \quad (3)$$

Una vez calculada la matriz de ofuscación M, se distribuye a los usuarios muestreados, quienes luego perturban su verdadera ubicación de acuerdo con las probabilidades codificadas en M e informan la ubicación perturbada al servidor.

4.2. Estimación de la distribución de probabilidad

Después de recopilar datos de ubicación perturbada de los usuarios muestreados, el siguiente paso es estimar la distribución de los usuarios en función de estos datos. Para cada cuadrícula $g_i \in G$, sea $P(g_i)$ la probabilidad de que un usuario esté ubicado en g_i . Luego, en esta subsección, estimamos $P(g_i)$ para todo $g_i \in G$ a partir de los datos de ubicación perturbada g'_j muestreados. $g'_j \in G$ son los datos perturbados que el servidor recibe de un usuario muestreado. Dejemos que g'_j sea una explicación, usaremos g'_j. Para que the denota la ubicación perturbada y g_j para denota la ubicación verdadera. La probabilidad de que esta ubicación perturbada se genere aleatoriamente a partir de la ubicación verdadera $g_i \in G$ se puede calcular de la siguiente manera:

$$P(g_i | g'_j) = \frac{P(g_i)P(g'_j | g_i)}{P(g'_j)} = \frac{P(g_i)P(g'_j | g_i)}{\sum_{g_k \in G} P(g_k)P(g'_j | g_k)} = \frac{P(g_i)M[i, j]}{\sum_{g_k \in G} P(g_k)M[k, j]} \quad (4)$$

Tenga en cuenta que $P(g'_j | g_i) = M[i, j]$ según la definición de la matriz de ofuscación. Como no es posible calcular la probabilidad previa, $P(g_i)$, directamente a partir de la ecuación anterior, debemos aproximarla. Existen varios métodos para aproximar la probabilidad previa, incluida la inferencia variacional [34], la cadena de Markov Monte Carlo [35] y la propagación de expectativas [36].

En este artículo, utilizamos el algoritmo de Maximización de Expectativas (EM) [37] para estimar $P(g_i)$. El algoritmo EM es particularmente efectivo cuando la probabilidad está bien definida, lo que en nuestro caso corresponde a la matriz de ofuscación, M.

Dejemos que DB sea una bolsa de datos de ubicación perturbados de usuarios muestreados. El proceso EM para estimar $P(g_i)$ para todo $g_i \in G$ a partir de DB es el siguiente.

- Inicialización: el parámetro (es decir, probabilidad previa) se inicializa de la siguiente manera:

$$P^{(0)}(g_1) = P^{(0)}(g_2) = \cdots = P^{(0)}(g_m) = \frac{1}{m} \quad (5)$$

- Paso E: la probabilidad posterior se calcula en función de los parámetros actuales de la siguiente manera:

$$P(g_i | g'_j) = \frac{P^{(t)}(g_i)M[i, j]}{\sum_{g_k \in G} P^{(t)}(g_k)M[k, j]} \quad (6)$$

- Paso M: El parámetro se actualiza utilizando las probabilidades posteriores actuales calculadas en el paso E anterior:

$$P^{(t+1)}(g_i) = \frac{\sum_{g'_j \in DB} P(g_i | g'_j)}{|DB|} \quad (7)$$

Aquí, |DB| representa el número de datos en la base de datos. Después de actualizar las probabilidades anteriores, realizamos un paso de normalización para garantizar que la suma de todas las probabilidades anteriores sea igual a 1 de la siguiente manera:

$$P_{(t+1)}(g_i) = \frac{P_{(t+1)}(g_i)}{\sum_{g_k \in G} P_{(t+1)}(g_k)} \quad (8)$$

Los pasos E y M anteriores se iteran hasta que el parámetro converge a un valor estable o el número de iteraciones alcanza un umbral predefinido.

4.3. Computación de cuadrículas particionadas adaptativamente

En esta subsección, presentamos un método que divide de forma adaptativa las cuadrículas según la distribución de probabilidad (es decir, P(gi) para todo gi G) calculada en la fase anterior. Inicialmente, el método propuesto trata todas las grillas en G como un solo grupo de grillas y luego divide iterativamente este grupo de arriba hacia abajo usando un algoritmo codicioso.

Sea $GC_v = \{C_1, C_2, \cdots, C_{|GC_v|}\}$ representa un conjunto de clústeres de grilla después de la v-ésima partición. Supongamos que para cada Ck GCv, grid(Ck) G denota el conjunto de grids que pertenecen al cluster Ck . Sea n el número total de usuarios de los que el servidor recopila datos de ubicación.

Luego, el número esperado de usuarios ubicados en la red gi se calcula como $Cnt(g_i) = n \cdot P(g_i)$.

Además, sea MGC_v un $|GC_v| \times |GC_v|$ matriz de ofuscación, que satisface -Geo-Ind, construida sobre elementos en GCv usando la Ecuación (3). La distancia entre dos grupos necesarios para calcular MGCv se determina utilizando los centroides de las cuadrículas que pertenecen a cada grupo. Luego, suponiendo que los usuarios perturban su ubicación de acuerdo con las probabilidades codificadas en MGCv , el número esperado de datos de ubicación perturbados correspondientes a las grillas pertenecientes a CK que el servidor recibe de n usuarios se calcula de la siguiente manera:

$$Cntperto(C_k) = \sum_{C_j \in GC_v} \sum_{g_i \in cuadrícula(C_j)} Cnt(g_i) \times M[j,k] \quad (9)$$

Sea Clus() una función que toma una cuadrícula como entrada y genera el clúster al que pertenece esa cuadrícula. Suponiendo que los usuarios están distribuidos uniformemente en las redes dentro de cada clúster, el error esperado debido a Geo-Ind con GCv se calcula de la siguiente manera:

$$ErrGC_v = \sum_{g_i \in G} Cnt(g_i) - \frac{Cntpert(Clus(g_i))}{tamaño(Clus(g_i))} \quad (10)$$

Aquí, tamaño (Ck) denota el número de cuadrículas que pertenecen al grupo Ck .

Supongamos que en la siguiente (v + 1)-ésima partición, se selecciona Ch GCv para dividirlo en subgrupos. En este artículo, dividimos Ch en cuatro subgrupos de igual tamaño dividiendo la región asociada horizontal y verticalmente. Sea GCh v+1 el conjunto de grupos de cuadrículas recién obtenidos al subdividir Ch . Utilizando el método descrito anteriormente, podemos estimar de manera similar el error esperado, ErrGCh v+1 . El conjunto de grupos de cuadrículas para la iteración (v + 1) se determina de la siguiente manera: causado por Geo-Ind con GCh

$$GC_{v+1} = \underset{1 \le h \le |GC_v|}{\operatorname{argmax}} \; ErrGCh_{v+1} - ErrGC_v \quad (11)$$

En otras palabras, un grupo de cuadrícula Ch que proporciona la máxima ganancia de reducción de errores es seleccionado para particionar en la siguiente iteración.

El algoritmo 1 presenta un pseudocódigo para dividir grillas de forma adaptativa utilizando la distribución de probabilidad de los usuarios. El algoritmo toma como entrada un conjunto de cuadrículas, G, y distribuciones de probabilidad, P(g1), · · · , P(gm), y genera un conjunto de grupos de cuadrículas, GC. En la línea 1, GCcur se inicializa para contener un único grupo que incluye todas las cuadrículas en G. Luego, en la línea 3, ErrGCcur se calcula usando GCcur. Entre las líneas 4 y 12, el algoritmo identifica el grupo Ch GCcur que produce la máxima ganancia de reducción de errores. Este proceso se repite hasta que la ganancia de reducción de errores sea mayor que 0. Finalmente, el algoritmo devuelve GCcur.

Algoritmo 1: Pseudocódigo para la entrada de partición de grilla adaptativa: G = {g1, · · · , gm} y P(gi) para toda la salida gi G : un conjunto de clusters de grilla GC

```
Inicialice GCcur;
si bien es cierto
    ErrGCcur = EstimarErr(GCcur, MGCcur ) ;
    Idx = 0, Gananciamejor = −∞;
    para h = 1 a |GCcur| hacer
        GCh_canalla = PartitionGrid(GCcur, h) ;
        ErrGCh_canalla = EstimateErr(GCh cur, MGCh_canalla ) ;
        si (ErrGCh_canalla − ErrGCcur ) > Gainbest entonces
            idx = h;
            Gananciamejor = ErrGCh_canalla − ErrGCcur ;
        fin
    fin
    si Gainbest > 0 entonces
        GCcur = GCIdx_canalla ;
    de lo
    contrario romper
fin
final
retorno GCcur;
```

El método de partición adaptativa de la red propuesto en esta subsección se basa en la distribución de probabilidad de los usuarios estimada a partir de datos muestreados. Por lo tanto, al igual que con otros métodos basados en muestreo, existe la posibilidad de que los datos muestreados estén sesgados. Estos sesgos pueden dar lugar a que se utilice una distribución de usuarios no representativa para la partición de la red, lo que puede conducir a una partición ineficaz. Esto, a su vez, puede afectar negativamente la utilidad general de los datos de ubicación recopilados, porque es posible que las cuadrículas adaptativas no reflejen con precisión la verdadera densidad de usuarios. Para mitigar posibles sesgos e imprecisiones al capturar la distribución de usuarios en tiempo real, se pueden considerar variaciones espaciales en el proceso de muestreo. Un método eficaz es utilizar el muestreo estratificado [38], que implica dividir toda la región en subregiones separadas. Al garantizar que cada subregión esté representada proporcionalmente en la muestra, el muestreo estratificado ayuda a reducir el sesgo de muestreo y proporciona una estimación más precisa de la distribución de usuarios.

4.4. Recopilación de datos de ubicación perturbada de los usuarios restantes utilizando cuadrículas con

particiones adaptativas. Se calcula una nueva matriz de ofuscación, M, utilizando el conjunto de cuadrículas con particiones adaptativas, GC, calculado en la fase anterior. Las cuadrículas divididas de forma adaptativa permiten un proceso de ofuscación más preciso y relevante al capturar la naturaleza dinámica de la distribución de usuarios de manera más efectiva que las cuadrículas estáticas. Una vez calculada la nueva matriz de ofuscación , se distribuye a los usuarios restantes. Luego, estos usuarios utilizan la matriz actualizada para alterar sus datos de ubicación reales, garantizando que su privacidad se preserve de acuerdo con los principios de Geo-Ind. Los datos de ubicación ofuscados luego se envían de regreso al servidor, donde se integran con los datos recopilados previamente de los usuarios muestreados.

5. Experimentos

En esta sección, primero describimos la configuración experimental. Luego, discutimos los resultados experimentales.

5.1. Configuración

experimental En esta sección, describimos los experimentos que llevamos a cabo para evaluar el enfoque propuesto. Para nuestros experimentos, utilizamos el conjunto de datos de trayectorias de taxi de Oporto [39],

Consiste en trayectorias de taxis compuestas por una serie de coordenadas GPS registradas de 442 taxis que operan en la ciudad de Oporto, Portugal. Extrajimos aleatoriamente 50.000 datos de ubicación de estas trayectorias, de los cuales 10.000 se consideraron datos de ubicación de los usuarios muestreados. En el experimento, variamos el número de cuadrículas de 400 (es decir, cuadrículas de 20 por 20) a 10.000 (es decir, cuadrículas de 100 por 100). En los experimentos, se informan los resultados para las siguientes alternativas: el método de cuadrícula no adaptativa (NG) existente en [13] y el método de cuadrícula adaptativa (AG) introducido en este artículo. Usamos las siguientes métricas para la evaluación: • La

métrica a nivel de datos mide la similitud entre el conjunto de datos de ubicación real y el conjunto de datos de ubicación perturbada recopilados bajo -Geo-Ind. Para la evaluación a nivel de datos, utilizamos tanto el error de conteo promedio (ACE) como el error de densidad. El error de conteo promedio cuantifica la diferencia entre el número real de usuarios, numtrue(gi), y el número derivado del conjunto de datos perturbados, numpert(gi), para cada cuadrícula. Es calculado como

$$\text{Error de conteo promedio} = \sum_{1 \le yo \le m} \frac{|\text{numtrue}(gi) - \text{numerert}(gi)|}{\text{máx}(\text{númeroverdadero}(gi), 1)} \quad (12)$$

El error de densidad mide la diferencia entre la distribución de densidad real de los usuarios y la versión perturbada calculada a partir de los conjuntos de datos recopilados bajo -Geo- Ind. Este error se mide como

$$\text{Error de densidad} = \text{JSD}(D(OD), D(PD)) \quad (13)$$

Aquí, JSD() representa la divergencia de Jenson-Shannon entre dos distribuciones del conjunto de datos de ubicación original, D(OD) y del conjunto de datos de ubicación perturbada, D(PD).

• La métrica a nivel de aplicación evalúa la utilidad de los datos recopilados desde la perspectiva de las aplicaciones que los utilizan. Utilizamos el error de consulta de rango para esta métrica, una medida ampliamente reconocida para evaluar la efectividad de los datos de ubicación [14]. En el experimento, generamos una consulta de rango, QR, con una región aleatoria R, y comparamos el número de resultados del conjunto de datos de ubicación original, QR(OD), con los de los conjuntos de datos de ubicación perturbados , QR(PD). Se calcula como

$$\text{Error de consulta de rango} = \frac{|QR(OD) - QR(PD)|}{\max(QR(OD), 1)} \quad (14)$$

En los experimentos, generamos 200 consultas de rango e informamos el error promedio de consulta de rango.

En el experimento, se utilizaron varios presupuestos de privacidad ( ) que oscilaban entre 0,2 y 2,0. Un presupuesto de privacidad inferior a 2 suele considerarse aceptable en aplicaciones prácticas [14]. Implementamos tanto NG como AG usando Python 3.8 y todos los experimentos se realizaron en un entorno equipado con CPU Intel Xeon 5220R y 64 MB de memoria.

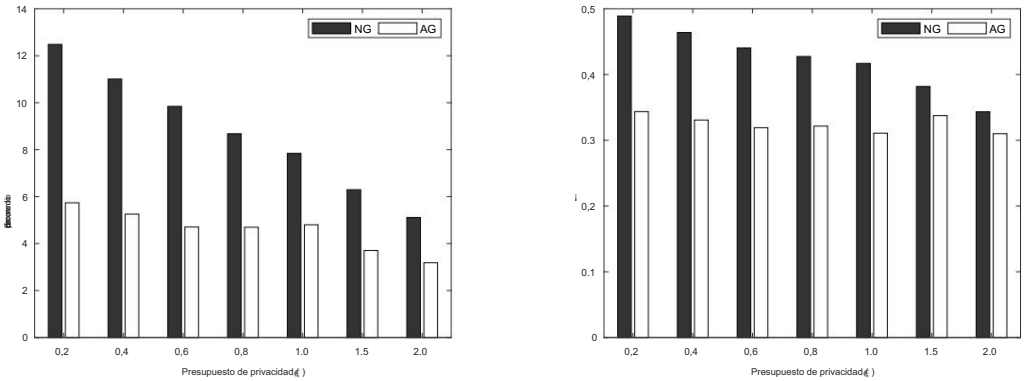
5.2. Resultados

En esta subsección, primero presentamos los resultados de la evaluación del nivel de datos y luego presentamos los resultados de la evaluación del nivel de aplicación.

5.2.1. Evaluación a nivel de datos

La Figura 2 muestra el efecto del presupuesto de privacidad tanto en el error de conteo promedio como en el error de densidad. En este experimento, el presupuesto de privacidad varía de 0,2 a 2,0, mientras que el tamaño de la cuadrícula se fija en 400. A medida que disminuye, ambos errores aumentan. Esto se debe a que a medida que disminuye, el grado de perturbación causada por Geo-Ind aumenta, lo que lleva a un mayor error, que se observa comúnmente con los métodos basados en DP. Como se muestra en las figuras, el método propuesto (AG) supera consistentemente al método existente (NG) en todos los niveles de presupuesto de privacidad. Además, la brecha de rendimiento entre los dos métodos aumenta a medida que aumenta la privacidad.

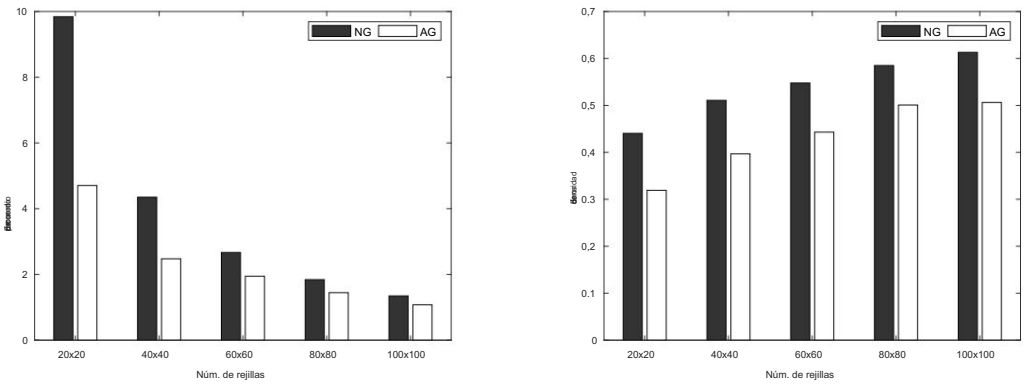
El presupuesto disminuye y, por tanto, aumenta el nivel de privacidad. Esto demuestra que la propuesta
El método es más ventajoso para aplicaciones que requieren un alto nivel de privacidad, lo cual es
típico de la mayoría de las aplicaciones que manejan datos de ubicación.



(a) (b)

Figura 2. Efecto del presupuesto de privacidad sobre (a) el error de conteo promedio y (b) el error de densidad para el método de red no adaptativa (NG) existente y el método de red adaptativa (AG) propuesto.

La Figura 3 muestra el efecto del número de cuadrículas en el error de conteo promedio y el
error de densidad. En este experimento, el número de rejillas varía de 400 a 10.000, mientras que el
El presupuesto de privacidad se fija en 0,6. La figura muestra que el método propuesto consistentemente
supera al método existente en todos los tamaños de cuadrícula. Más concretamente, como se puede ver en
En la figura, el error de conteo promedio disminuye a medida que aumenta el número de cuadrículas. Tenga en cuenta que
A medida que aumenta el número de redes, el número de usuarios por red disminuye porque el total
El número de usuarios es fijo. Esto, a su vez, reduce el error de conteo promedio, que se basa en
la diferencia absoluta entre los usuarios obtenidos a partir de los datos de ubicación reales y los
datos de ubicación perturbados. Por otro lado, a medida que aumenta el número de rejillas, la densidad
error, que mide la divergencia de Jenson-Shannon entre dos distribuciones de la
El conjunto de datos de ubicación original y los conjuntos de datos de ubicación perturbados aumentan. Esto se debe a que el
La divergencia de Jenson-Shannon mide la diferencia relativa entre distribuciones y
por lo tanto, no se ve afectado por el número de usuarios por red. A medida que el tamaño de la cuadrícula se vuelve más fino,
las perturbaciones en los datos tienen un efecto más pronunciado en la distribución, lo que resulta
en una mayor divergencia entre los conjuntos de datos originales y perturbados.



(a) (b)

Figura 3. Efecto del número de cuadrículas sobre (a) el error de conteo promedio y (b) el error de densidad para el método de red no adaptativa (NG) existente y el método de red adaptativa (AG) propuesto.

Los resultados mostrados en las Figuras 2 y 3 confirman que el método propuesto permite la
recopilación de datos de ubicación que son más similares a los datos originales bajo Geo-Ind que
el método existente. Estos resultados resaltan las importantes ventajas de nuestro enfoque.

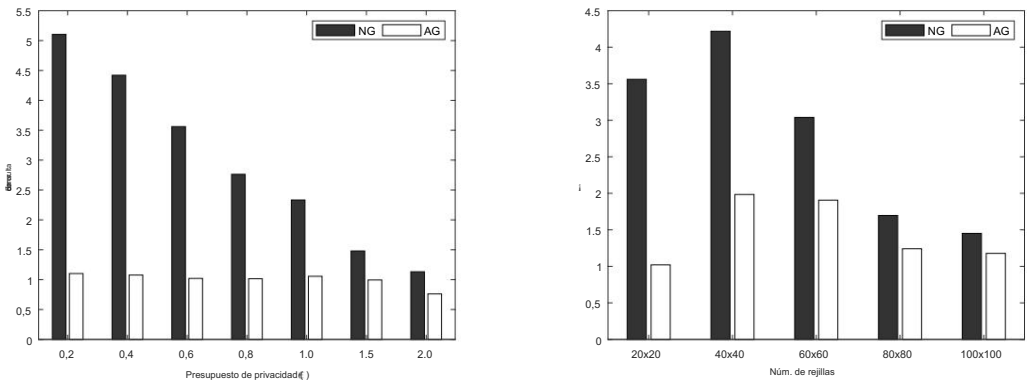
en la recopilación de datos de ubicación que preservan la privacidad. Al ajustar dinámicamente las cuadrículas en función de la distribución de usuarios en tiempo real bajo Geo-Ind, logramos una mayor precisión y utilidad de los datos.

5.2.2. Evaluación a nivel de aplicación La

Figura 4 ilustra el efecto del presupuesto de privacidad y el número de cuadrículas en el error de consulta de rango. En la Figura 4a, el tamaño de la cuadrícula se fija en 400, mientras que en la Figura 4b, el presupuesto de privacidad se fija en 0,6. Como se muestra en la figura, a medida que disminuye, el error asociado con el método existente aumenta drásticamente, mientras que el error del método propuesto aumenta sólo marginalmente. Esto ocurre porque el método propuesto es capaz de recopilar conjuntos de datos de ubicación que están más cerca de los conjuntos de datos originales, como lo verifica la evaluación a nivel de datos.
La solidez de nuestro enfoque bajo diferentes presupuestos de privacidad resalta su efectividad para equilibrar la privacidad y la precisión. A medida que se vuelve más pequeño, lo que indica mayores garantías de privacidad, el método propuesto aún logra preservar la utilidad de los datos, haciéndolos más confiables para aplicaciones que requieren información de ubicación precisa.

Además, el método propuesto supera consistentemente al método existente en todos los tamaños de cuadrícula. Esto verifica que nuestro método propuesto es robusto independientemente del número de cuadrículas. La capacidad de mantener bajas tasas de error de consulta en diferentes configuraciones de red demuestra la adaptabilidad y eficacia de nuestro enfoque. Esta solidez es crucial para aplicaciones prácticas que requieren diferente granularidad en la representación de la ubicación del usuario (es decir, número de cuadrículas) según los requisitos de la aplicación.

Estos resultados experimentales indican que el método propuesto se puede utilizar para una amplia gama de servicios y aplicaciones basados en la ubicación que requieren diferentes niveles de privacidad y granularidad en la representación de la ubicación del usuario.



(a) (b)

Figura 4. Efecto de (a) el presupuesto de privacidad y (b) el número de cuadrículas en el error de consulta de rango para el método de cuadrícula no adaptativa (NG) existente y el método de cuadrícula adaptativa (AG) propuesto.

5.2.3. Evaluación de los efectos de la variabilidad de la red y la adaptación de la red

Todos los resultados de los experimentos de las subsecciones anteriores se obtuvieron bajo el supuesto de que las condiciones de la red son estables. Sin embargo, en escenarios del mundo real, las condiciones de la red pueden ser muy variables e impredecibles. El método propuesto divide de forma adaptativa la cuadrícula en función de los datos de ubicación muestreados recopilados durante el proceso de recopilación de datos de ubicación. Sin embargo, las condiciones inestables de la red, como una alta latencia de la red, la pérdida de paquetes y un ancho de banda bajo, pueden retrasar la recopilación oportuna de estos datos muestreados. Como resultado, es posible que algunos datos de muestra no estén disponibles para el cálculo de la partición de la red adaptativa, lo que puede dar lugar a una partición de la red menos precisa.

En esta subsección, para abordar los desafíos de las redes inestables, evaluamos la efectividad del método propuesto en escenarios de redes reales. El experimento que se muestra en la Tabla 1 considera un escenario en el que algunos de los datos de ubicación muestreados necesarios para estimar la distribución de usuarios en la Sección 4.2 (que luego se utiliza para calcular de forma adaptativa las particiones de la red en la Sección 4.3) se pierden o no se reciben a tiempo debido a red inestable

condiciones tales como alta latencia, pérdida de paquetes y bajo ancho de banda. En este experimento, el La tasa de pérdida de datos de ubicación muestreados varía del 1% al 20%, cubriendo un rango desde típico a condiciones severas de la red. En el experimento, el número de cuadrículas se establece en 400 y el presupuesto de privacidad se establece en 0,6. Como se muestra en la Tabla 1, incluso cuando la tasa de pérdida de muestras Los datos de ubicación aumentan del 1% al 20%, ambos errores permanecen estables con solo una muy pequeña aumentar. En particular, incluso en condiciones de red severas con una tasa de pérdida del 20%, el El método de red adaptativa (AG) propuesto supera significativamente a la red no adaptativa (NG) método. Estos resultados verifican la robustez y eficacia del método propuesto en mantener la utilidad de los datos en escenarios de red reales con diferentes niveles de pérdida de datos debido a condiciones de red inestables.

Tabla 1. Efecto de la tasa de pérdida de datos de ubicación muestreados sobre el error de conteo promedio y el error de densidad.

	AG				NG
Tasa de pérdida de datos de ubicación muestreados 0%	1%	5%	10%	20%	
Error de recuento promedio	4.705	4.867	4.896	4.936	5.018	9.842
Error de densidad	0.319	0.321	0.328	0.336	0.341	0.440

El método propuesto ajusta de forma adaptativa la cuadrícula durante el proceso de recopilación de datos de ubicación, lo que introduce latencia adicional debido a la sobrecarga computacional de calcular la red adaptativa. Por lo tanto, evaluamos experimentalmente la latencia introducida por el método propuesto. La Tabla 2 muestra los resultados de latencia causados por la grilla adaptativa. cálculo del método propuesto. En este experimento, el número de rejillas varía de 400 a 10.000, mientras que el presupuesto de privacidad se fija en 0,6. Como se muestra en la tabla, la latencia aumenta a medida que aumenta el tamaño de la cuadrícula. Esto se debe a que a medida que aumenta el número de cuadrículas, la El número de iteraciones necesarias para particionar adaptativamente las cuadrículas también aumenta. Como resultado, Las redes más grandes requieren más recursos computacionales, lo que conduce a una mayor latencia.

Tabla 2. Latencia causada por el cálculo de la grilla adaptativa.

Tamaño de la cuadrícula	20 × 20	40 × 40	60 × 60	80 × 80	100 × 100
Tiempo(s) de ejecución	4.10	17.41	43.12	139,56	439.56

Observamos que aunque el cálculo de la cuadrícula adaptativa introduce latencia adicional como se muestra en la Tabla 2, es un proceso único dentro de la recopilación general de datos de ubicación. procedimiento. Por lo tanto, el impacto de estos gastos generales en el tiempo total de procesamiento del La recopilación de datos de ubicación es limitada. Además, la latencia adicional causada por la sobrecarga computacional de la partición adaptativa de la red puede mitigarse mediante varios procesos paralelos. técnicas de procesamiento. En particular, técnicas como los marcos de computación distribuida [40,41] y la aceleración de GPU pueden reducir significativamente el tiempo de cálculo, con lo que mitigar esta latencia. Al distribuir la carga de trabajo entre múltiples procesadores, estos Los enfoques pueden mejorar la eficiencia del proceso de partición adaptativa de la red, asegurando Análisis de datos en aplicaciones en tiempo real.

6. Conclusiones y trabajo futuro

Recientemente, ha habido una demanda creciente de recopilar y compartir información datos de localización. Dada la naturaleza sensible de la información sobre la ubicación del usuario, se requieren esfuerzos considerables Se han tomado medidas para garantizar la privacidad, surgiendo esquemas diferenciales basados en la privacidad como el enfoque preferido. Sin embargo, estos esquemas normalmente representan ubicaciones de usuarios en cuadrículas divididas uniformemente, que a menudo no reflejan con precisión la verdadera distribución de usuarios dentro de un espacio. En este artículo, presentamos un enfoque novedoso que ajusta dinámicamente la cuadrícula en tiempo real durante la recopilación de datos de ubicación utilizando Geo-Ind para mejorar la utilidad de los datos recopilados. El método propuesto captura la distribución de usuarios directamente durante los datos. recopilación, eliminando la dependencia de información de distribución preexistente. Experimental

Los resultados sobre datos reales confirmaron que el esquema propuesto mejora significativamente la utilidad de los datos de ubicación recopilados tanto a nivel de datos como de aplicación. Específicamente, los resultados mostraron que, en comparación con la solución existente, el método propuesto puede reducir la tasa de error hasta en un 52 % en experimentos a nivel de datos y hasta en un 75 % en experimentos a nivel de aplicación.

A pesar de los resultados prometedores, el método propuesto tiene las siguientes limitaciones. Dado que el método propuesto ajusta de forma adaptativa la cuadrícula en tiempo real durante la recopilación de datos, existe una sobrecarga computacional adicional asociada con el cálculo de la cuadrícula adaptativa. Esta sobrecarga es especialmente significativa cuando se utilizan cuadrículas de gran tamaño. Por lo tanto, el trabajo futuro se centrará en mejorar la eficiencia del cálculo de la red adaptativa, especialmente para redes de gran tamaño con una gran cantidad de usuarios. Esto se puede lograr al paralelizar el proceso de partición de la red adaptativa para reducir la sobrecarga computacional. Exploraremos varias técnicas de procesamiento paralelo, como la implementación de marcos informáticos distribuidos como Apache Hadoop [40] o Apache Spark [41], que distribuyen datos y cálculos a través de un grupo de máquinas. Además, se puede considerar el uso de subprocesos múltiples dentro de una sola máquina y el uso de aceleración de GPU para aumentar la eficiencia. Además, la integración de servicios de computación en la nube mejorará la escalabilidad al proporcionar una infraestructura dinámica y escalable para realizar particiones de red adaptativas en grandes conjuntos de da

Otra dirección de investigación futura es analizar teóricamente el impacto de la partición adaptativa de la red en la utilidad de los datos de ubicación recopilados. Este análisis aclarará los principios subyacentes de la partición adaptativa de la red y su impacto en la utilidad de los datos, proporcionando así un apoyo teórico más sólido para el método propuesto. Además, se puede investigar más a fondo el equilibrio entre privacidad y utilidad para optimizar el equilibrio entre privacidad y utilidad. Esto incluirá el desarrollo de modelos y métricas para evaluar cuantitativamente este equilibrio en diversas condiciones.

Financiamiento: Esta investigación fue financiada por una Beca de Investigación 2023 de la Universidad Sangmyung (2023-A000-0119).

Declaración de disponibilidad de datos: los datos originales presentados en el estudio están disponibles abiertamente en Kaggle en https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i (consultado el 20 de junio de 2024).

Conflictos de intereses: El autor declara no tener conflictos de intereses.

Abreviaturas

En este manuscrito se utilizan las siguientes abreviaturas:

PD	Privacidad diferencial
PLD	Privacidad diferencial local
Privacidad diferencial de métricas de MDP	
Geo-Ind Geo-Indistinguibilidad	
MCS	Crowdsensing móvil
EM	Maximización de expectativas
JSD	Divergencia Jenson-Shannon

Referencias

1. Wang, X.; Puede.; Wang, Y.; Jin, W.; Wang, X.; Tang, J.; Jia, C.; Yu, J. Predicción del flujo de tráfico mediante gráfico temporal espacial neuronal red. En Actas de la conferencia web, Taipei, Taiwán, 20 a 24 de abril de 2020; págs. 1082-1092.
2. Pan, Z.; Liang, Y.; Wang, W.; Yu, Y.; Zheng, Y.; Zhang, J. Predicción del tráfico urbano a partir de datos espacio-temporales mediante metaaprendizaje profundo . En actas de la Conferencia internacional ACM SIGKDD sobre descubrimiento de conocimientos y minería de datos, Anchorage, Alaska, EE. UU., 4 a 8 de abril de 2019; págs. 1720-1730.
3. Kim, JS; Kim, JW; Chung, YD Recomendación sucesiva de puntos de interés con privacidad diferencial local. Acceso IEEE 2021, 9, 66371–66386. [Referencia cruzada]
4. An, J.; Li, G.; Jiang, W. NRDL: Aprendizaje descentralizado de las preferencias del usuario para la próxima recomendación de PDI que preserve la privacidad. Experto Sistema. Aplica. 2024, 239, 122421. [Referencia cruzada]
5. Primault, V.; Boutet, A.; Mokhtar, SB; Brunie, L. El largo camino hacia la privacidad de la ubicación computacional: una encuesta. Comunicaciones IEEE. Sobrevivir. Tutor. 2019, 21, 2772–2793. [Referencia cruzada]

6. Liu, B.; Zhou, W.; Zhu, T.; Gao, L.; Xiang, Y. Privacidad de la ubicación y sus aplicaciones: un estudio sistemático. Acceso IEEE 2019, 6, 17606–17624. [Referencia cruzada]

7. Alharthi, R.; Banihani, A.; Alzahrani, A.; Alshehri, A.; Alshahrani, H.; Fu, H.; Liu, A.; Zhu, Y. Desafíos de la privacidad de la ubicación en el crowdsourcing espacial. En Actas de la Conferencia Internacional IEEE sobre Tecnología Electrónica/Información, Rochester, MI, EE. UU., 3 al 5 de mayo de 2018.

8. Henriksen-Bulmer, J.; Jeary, S. Ataques de reidentificación: una revisión sistemática de la literatura. En t. J.Inf. Gestionar. 2016, 36, 1184–1192. [Referencia cruzada]

9. Dwork, C. Privacidad diferencial. En Actas del Coloquio Internacional sobre Autómatas, Lenguajes y Programación, Venecia, Italia, 10 a 14 de julio de 2006; págs. 1–12.

10. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibración del ruido a la sensibilidad en el análisis de datos privados. En Actas de la Tercera conferencia sobre Teoría de la Criptografía, Nueva York, NY, EE.UU., 4 a 7 de marzo de 2006.

11. Bordenabe, NE; Chatzikokolakis, K.; Palamidess, C. Mecanismos geoindistinguibles óptimos para la privacidad de la ubicación. En Actas de la Conferencia ACM SIGSAC sobre Seguridad Informática y de las Comunicaciones, Nueva York, NY, EE. UU., 3 a 7 de noviembre de 2014; págs. 251–262.

12. Kim, J.; Jang, B. Recopilación de datos de posicionamiento en interiores consciente de la carga de trabajo a través de privacidad diferencial local. Comunicaciones IEEE. Letón. 2019, 23, 1352-1356. [Referencia cruzada]

13. Zhang, P.; Cheng, X.; Su, S.; Wang, N. Reclutamiento de trabajadores basado en la cobertura del área bajo geoindistinguibilidad. Computadora. Neto. 2022, 217, 109340. [Referencia cruzada]

14. Du, Y.; Hu, Y.; Zhang, Z.; Colmillo, Z.; Chen, L.; Zheng, B.; Gao, Y. LDPTrace: Síntesis de trayectoria privada localmente diferencial. En Actas del VLDB Endowment, Vancouver, BC, Canadá, 28 de agosto a 1 de septiembre de 2023; págs. 1897-1909.

15. Ghaemi, Z.; Farnaghi, M. Un enfoque de agrupación basado en densidad variada para la detección de eventos a partir de datos heterogéneos de Twitter. ISPRS Int. J. Geo-Inf. 2019, 8, 82. [Referencia cruzada]

16. Alvim, M.; Chatzikokolakis, K.; Palamidessi, C.; Pazii, A. Privacidad diferencial local en espacios métricos: optimización del equilibrio con la utilidad. En Actas del Simposio IEEE Computer Security Foundations, Oxford, Reino Unido, 9 a 12 de julio de 2018.

17. Andrés, YO; Bordenabe, NE; Chatzikokolakis, K.; Palamidessi, C. Geoindistinguibilidad: privacidad diferencial para sistemas basados en la ubicación. En Actas de la Conferencia ACM SIGSAC sobre Seguridad Informática y de las Comunicaciones, Berlín, Alemania, 4 a 8 de noviembre de 2013; págs. 901–914.

18. Chatzikokolakis, K.; Palamidessi, C.; Stronati, M. Geoindistinguibilidad: un enfoque basado en principios para la privacidad de la ubicación. En Actas de la Conferencia Internacional sobre Computación Distribuida y Tecnología de Internet, Bhubaneswar, India, 5 a 8 de febrero de 2015; págs. 49–72.

19. Wang, L.; Yang, D.; Han, X.; Wang, T.; Zhang, D.; Ma, X. Asignación de tareas que preservan la privacidad de la ubicación para detección de multitudes móvil con ofuscación geográfica diferencial. En Actas de la Conferencia Internacional sobre la World Wide Web, Perth, Australia, 3 a 7 de abril de 2017; págs. 627–636.

20. Qiu, C.; Squicciarini, AC Protección de la privacidad de la ubicación en el crowdsourcing espacial basado en vehículos a través de la geoindistinguibilidad. En Actas de la Conferencia Internacional IEEE sobre Sistemas de Computación Distribuida, Dallas, TX, EE. UU., 7 a 10 de julio de 2019; págs. 1061-1071.

21. Jin, W.; Xiao, M.; Guo, L.; Yang, L.; Li, M. ULPT: Un marco de comercio de privacidad de ubicación centrado en el usuario para la detección de multitudes en dispositivos móviles. Traducción IEEE. Multitud. Computadora. 2022, 21, 3789–3806. [Referencia cruzada]

22. Huang, P.; Zhang, X.; Guo, L.; Li, M. Incentivar el monitoreo del ruido basado en crowdsensing con ubicaciones diferencialmente privadas. Traducción IEEE. Multitud. Computadora. 2021, 20, 519–532. [Referencia cruzada]

23. Zhao, Y.; Yuan, D.; Du, JT; Chen, J. Geo-Ellipse-Indistinguibilidad: protección de privacidad de ubicación consciente de la comunidad para direccionales distribución. Traducción IEEE. Conocimiento. Ing. de datos. 2023, 35, 6957–6967. [Referencia cruzada]

24. Yu, L.; Zhang, S.; Meng, Y.; Du, S.; Chen, Y.; Ren, Y.; Zhu, H. Publicidad basada en la ubicación que preserva la privacidad mediante la geoindistinguibilidad longitudinal. Traducción IEEE. Multitud. Computadora. 2024, 23, 8256–8273. [Referencia cruzada]

25. Zhao, Y.; Chen, J. Indistinguibilidad de vectores: protección de la privacidad basada en la dependencia de la ubicación para datos de ubicación sucesivos. IEEE Trans. Computadora. 2024, 73, 970–979. [Referencia cruzada]

26. Mendes, R.; Cunha, M.; Vilela, JP Geo-indistinguibilidad consciente de la velocidad. En actas de la Conferencia ACM sobre datos y Seguridad y privacidad de aplicaciones, Charlotte, Carolina del Norte, EE. UU., 24 a 26 de abril de 2023; págs. 141-152.

27. Ren, W.; Tang, S. EGeoIndis: Un marco de protección de la privacidad de la ubicación eficaz y eficiente en la detección de la densidad del tráfico. Veh. Comunitario. 2020, 21, 100187. [Referencia cruzada]

28. Kim, JW; Lim, B. Estimación efectiva y que preserva la privacidad de la distribución de densidad de usuarios de LBS bajo geoindistinguibilidad . Electrónica 2023, 12, 917. [CrossRef]

29. Chen, R.; Pequeño.; Puede.; Gong, Y.; Guo, Y.; Ohtsuki, T.; Pan, M. Construcción de un mapa de vulnerabilidad COVID-19 de colaboración colectiva para dispositivos móviles Con Geo-Indistinguibilidad. IEEE Internet Things J. 2022, 9, 17403–17416. [Referencia cruzada]

30. Fathalizadeh, A.; Moghtadaiee, V.; Alishahi, M. Geoindistinguibilidad en interiores: adopción de privacidad diferencial para interiores Protección de datos de ubicación. Traducción IEEE. Emergente. Arriba. Computadora. 2023, 12, 293–306. [Referencia cruzada]

31. Feyisetan, O.; Balle, B.; Drake, T.; Diethe, T. Análisis textual que preserva la privacidad y la utilidad mediante perturbaciones multivariadas calibradas . En Actas de la Conferencia Internacional sobre Búsqueda Web y Minería de Datos, Houston, TX, EE. UU., 3 a 7 de febrero de 2020; págs. 178–186.

32. Canción, S.; Kim, JW Adaptación de la geoindistinguibilidad para la recopilación de microdatos médicos que preservan la privacidad. Electrónica 2023, 12, 2793. [Referencia cruzada]
33. Ahuja, R.; Ghinita, G.; Shahabi, C. Una técnica escalable y que preserva la utilidad para proteger los datos de ubicación con geoindistinguibilidad . En Actas de la Conferencia Internacional sobre la Ampliación de la Tecnología de Bases de Datos, Lisboa, Portugal, 26 a 29 de marzo de 2019; págs. 210-231.
34. Blei, DM; Kucukelbir, A.; McAuliffe, JD Inferencia variacional: una revisión para estadísticos. Mermelada. Estadística. Asociación 2017, 112, 859–877. [Referencia cruzada]
35. Chib, S. Métodos de Monte Carlo de la cadena de Markov: cálculo e inferencia. Mano. Economía. 2001, 5, 3569–3649.
36. Li, Y.; Hernández-Lobato, JM; Turner, RE Propagación de expectativas estocásticas. arXiv 2018, arXiv:1506.04132.
37. Sammaknejad, N.; Zhao, Y.; Huang, B. Una revisión del algoritmo de maximización de expectativas en la identificación de procesos basada en datos. J. Control de procesos 2019, 73, 123–136. [Referencia cruzada]
38. Howell, CR; Su, W.; Nassel, AF; Agné, AA; Cherrington, AL Muestreo aleatorio estratificado basado en áreas utilizando datos geoespaciales tecnología en una encuesta comunitaria. BMC Public Health 2020, 20, 1678. [CrossRef] [PubMed]
39. Moreira-Matías, L.; Gama, J.; Ferreira, M.; Mendes-Moreira, J.; Damas, L. Predicción de la demanda de taxis y pasajeros mediante streaming datos. Traducción IEEE. Intel. Transp. Sistema. 2013, 14, 1393–1402. [Referencia cruzada]
40. Apache Hadoop. Disponible en línea: https://hadoop.apache.org/ (consultado el 22 de julio de 2024).
41. Apache Spark: motor unificado para análisis de datos a gran escala. Disponible en línea: https://spark.apache.org/ (accedido en 22 de julio de 2024).