

Robustez y generalización de modelos de aprendizaje automático y profundo en la clasificación de perturbaciones de calidad de energía

Robustness and Generalization of Machine Learning and Deep Learning Models for Power Quality Disturbance Classification

Recibido: 14 de enero de 2026

Aprobado: 21 de mayo de 2026

Cómo citar: D. O. Cardozo Sarmiento, K. C. Puerto López, P. A. Gómez Monsalve, and D. Guevara Ibarra, "Robustez y generalización de modelos de aprendizaje automático y profundo en la clasificación de perturbaciones de calidad de energía," *Mundo Fesc*, vol. 16, no. 36, pp. 51-74, Sep. 2026, doi: 10.61799/2216-0388.2373.

Darwin Orlando Cardozo Sarmiento^{1*}



Doctor en Ingeniería Eléctrica y Electrónica
darwinorlandocs@ufps.edu.co
<https://orcid.org/0000-0003-3177-3893>
Universidad Francisco de Paula Santander, Cúcuta, Colombia

Karla Cecilia Puerto López²



Doctor en Educación
karlaceciliapl@ufps.edu.co
<https://orcid.org/0000-0003-3749-676X>
Universidad Francisco de Paula Santander, Cúcuta, Colombia

Pablo Andrés Gómez Monsalve³



Doctor en Automática
pablo.gomez@unipamplona.edu.co
<https://orcid.org/0000-0002-9507-1690>
Universidad de Pamplona, Pamplona, Colombia

Dinael Guevara Ibarra⁴



Doctor en Ingeniería
dinaelgi@ufps.edu.co
<https://orcid.org/0000-0003-3007-8354>
Universidad Francisco de Paula Santander, Cúcuta, Colombia

***Autor para correspondencia:**

darwinorlandocs@ufps.edu.co



Robustez y generalización de modelos de aprendizaje automático y profundo en la clasificación de perturbaciones de calidad de energía

Palabras clave: Calidad de energía; perturbaciones eléctricas; aprendizaje automático; aprendizaje profundo; generalización entre dominios; robustez; AWGN.

Resumen

La clasificación automática de perturbaciones de calidad de energía suele evaluarse bajo condiciones in-domain, lo que puede sobreestimar el desempeño cuando cambian la distribución de los datos o las condiciones de ruido. Este estudio evaluó la robustez y la capacidad de generalización de modelos de aprendizaje automático y profundo para la clasificación de ocho tipos de perturbaciones eléctricas. Se armonizaron tres dominios de datos —SEED-PQD, PQD-16c y un conjunto sintético— a una representación común de un ciclo, 64 muestras y 3,2 kHz. Se compararon máquinas de vectores de soporte (SVM), Random Forest, XGBoost, 1D-CNN y CNN-LSTM mediante Macro-F1, exactitud balanceada, MCC, generalización cross-domain y degradación frente a ruido blanco gaussiano aditivo entre 30 y 0 dB. Los modelos profundos alcanzaron Macro-F1 in-domain superiores a 0,98 en SEED y en el dominio sintético cuando dispusieron de suficientes datos, aunque mostraron elevada sensibilidad al tamaño del conjunto de entrenamiento y al ruido. SVM registró el mayor Macro-F1 cross-domain promedio (0,6407) y los valores más altos de AUC de robustez (0,4328) y AUC de retención (0,5039). La prueba de Friedman confirmó diferencias significativas entre los modelos bajo ruido ($p < 0,001$; $W = 0,5265$), mientras que en generalización cross-domain no se detectaron diferencias globales significativas ($p = 0,1257$). Los resultados muestran que un elevado desempeño in-domain no garantiza generalización ni robustez, y evidencian la necesidad de evaluar conjuntamente precisión, transferencia entre dominios y tolerancia al ruido antes de seleccionar modelos para aplicaciones reales de calidad de energía.

Robustness and Generalization of Machine Learning and Deep Learning Models for Power Quality Disturbance Classification

Keywords: Power quality; power quality disturbances; machine learning; deep learning; cross-domain generalization; robustness; AWGN.

Abstract

Automatic power quality disturbance classification is commonly evaluated under in-domain conditions, which may overestimate performance when data distributions or noise conditions change. This study evaluated the robustness and generalization capability of machine learning and deep learning models for the classification of eight power quality disturbance classes. Three data domains—SEED-PQD, PQD-16c, and a synthetic dataset—were harmonized to a common representation of one cycle, 64 samples, and a 3.2 kHz sampling rate. Support vector machines (SVM), Random Forest, XGBoost, 1D-CNN, and CNN-LSTM were compared using Macro-F1, balanced accuracy, MCC, cross-domain generalization, and degradation under additive white Gaussian noise from 30 to 0 dB. Deep learning models achieved in-domain Macro-F1 values above 0.98 on SEED and the synthetic domain when sufficient training data were available, although they showed high sensitivity to training-set size and noise. SVM achieved the highest average cross-domain Macro-F1 (0.6407) and the highest robustness AUC (0.4328) and retention AUC (0.5039). The Friedman test confirmed significant differences among models under noise ($p < 0.001$; $W = 0.5265$), whereas no statistically significant global differences were detected for cross-domain generalization ($p = 0.1257$). The results show that high in-domain performance does not guarantee generalization or robustness and highlight the need to jointly evaluate classification performance, domain transfer, and noise tolerance before selecting models for real-world power quality applications.

Introducción

La calidad de la energía eléctrica constituye un aspecto fundamental para la operación confiable de los sistemas eléctricos modernos. La incorporación creciente de fuentes renovables, cargas no lineales, convertidores electrónicos de potencia y equipos sensibles ha incrementado la presencia y diversidad de perturbaciones capaces de modificar la forma de onda de tensión y corriente. Entre estas perturbaciones se encuentran los hundimientos y elevaciones de tensión, interrupciones, armónicos, flicker, transitorios oscilatorios y muescas, cuya identificación oportuna es necesaria para apoyar procesos de supervisión, diagnóstico y toma de decisiones en redes eléctricas. En este contexto, la clasificación automática de perturbaciones de calidad de energía (Power Quality Disturbances, PQD) se ha convertido en un área activa de investigación, en la que se combinan técnicas de procesamiento de señales con métodos de inteligencia artificial [1], [2]. Las revisiones recientes muestran que los métodos basados en aprendizaje automático y aprendizaje profundo pueden alcanzar elevados niveles de clasificación, aunque persisten problemas relacionados con disponibilidad de datos, sobreajuste, interpretabilidad y adaptación a condiciones diferentes de aquellas utilizadas durante el entrenamiento [1], [2].

Enfoques tradicionales de aprendizaje automático como vectores de soporte (SVM), bosques aleatorios (RF) y métodos de boosting, permiten construir clasificadores a partir de características estadísticas, temporales y espectrales extraídas de las señales. El aprendizaje profundo favorece el desarrollo de modelos que aprenden directamente representaciones a partir de las formas de onda. Las redes convolucionales unidimensionales (1D-CNN) pueden capturar patrones locales y transitorios y arquitecturas híbridas CNN-LSTM incorporan mecanismos para representar características locales y dependencias secuenciales. También se han propuesto representaciones tiempo-frecuencia y esquemas de transferencia de aprendizaje para mejorar el reconocimiento de diferentes tipos de perturbaciones [1]-[3]. Sin embargo, la mayoría de los estudios se concentra en maximizar la exactitud dentro del mismo conjunto de datos utilizado para desarrollar el modelo, que puede no representar su comportamiento ante distribuciones diferentes.

Esta limitación en un clasificador con elevado desempeño in-domain puede experimentar una degradación considerable cuando cambian características estadísticas, procedimiento de adquisición, frecuencia de muestreo, generación de las señales o condiciones operativas. La literatura ha comenzado a abordar este problema con transferencia de aprendizaje, clasificación de perturbaciones desconocidas y evaluaciones entre dominios. Se han propuesto esquemas de transferencia para actualizar clasificadores CNN-LSTM frente a nuevas perturbaciones y cambios de distribución, también se han incorporado pruebas cross-domain sobre datos sintéticos y conjuntos reales como SEED [4]-[6]. Estos avances demuestran que la capacidad de generalización debe evaluarse

de forma independiente del rendimiento obtenido sobre datos procedentes de la misma distribución de entrenamiento.

La robustez frente al ruido representa otro desafío; las señales adquiridas en sistemas eléctricos reales pueden contener ruido de medición, interferencias y perturbaciones no consideradas durante el entrenamiento. Estudios han demostrado que modelos con alta exactitud en condiciones nominales pueden reducir su desempeño cuando se modifican las entradas, ya sea por ruido convencional o perturbaciones adversariales. Para este desafío, se han desarrollado estrategias de clasificación que evalúan el comportamiento con diferentes relaciones señal-ruido y arquitecturas orientadas a mejorar la tolerancia al ruido. La evaluación comparativa de modelos de aprendizaje automático y profundo mostró la importancia de incorporar ruido gaussiano y niveles de relación señal-ruido en los experimentos [7]-[10].

La cantidad de información disponible para el entrenamiento también representa otro desafío. Las redes profundas tienen gran capacidad de representación, pero requieren un volumen suficiente de datos para aprender características generalizables; la escasez de registros puede producir sobreajuste o una elevada dependencia de la inicialización. Gran parte de la literatura sobre calidad de energía utiliza datos sintéticos y bases de datos abiertas muy limitadas y diversas para establecer comparaciones homogéneas [1]-[2]. Por esta razón, comparar algoritmos mediante solo exactitud puede conducir a conclusiones incompletas: la selección de un modelo para una aplicación práctica debería considerar su desempeño nominal, requerimiento de datos, estabilidad, generalización hacia otros dominios, robustez frente al ruido y costo computacional.

Los estudios que analizan bajo un mismo protocolo experimental modelos clásicos y profundos utilizando varios dominios de datos, clases comunes, evaluación in-domain, transferencia directa cross-domain, sensibilidad a la disponibilidad de datos y degradación controlada frente al ruido son muy limitados. El crecimiento de estudios enfocados en arquitecturas más complejas dificulta establecer si mayor exactitud dentro de un conjunto de datos implica mayor capacidad de transferencia o robustez. Este trabajo aborda este vacío mediante la armonización de tres dominios independientes de señales de calidad de energía y la comparación de SVM, Random Forest, XGBoost, 1D-CNN y CNN-LSTM sobre ocho clases comunes de perturbaciones.

El objetivo de esta investigación es evaluar comparativamente la robustez y la capacidad de generalización de modelos de aprendizaje automático y aprendizaje profundo para la clasificación de perturbaciones de calidad de energía, mediante pruebas in-domain y cross-domain, análisis de suficiencia de datos y evaluación frente a ruido blanco gaussiano aditivo (AWGN) bajo diferentes relaciones señal-ruido. El análisis se complementa mediante métricas de clasificación, medidas de degradación y retención del desempeño, pruebas estadísticas y criterios de costo computacional, con el propósito de determinar hasta qué punto el rendimiento observado en condiciones controladas se conserva cuando cambian las características del dominio y la calidad de las señales.

Materiales y Métodos

Diseño y contexto del estudio

Se desarrolla un estudio cuantitativo de tipo experimental-computacional y comparativo orientado a evaluar el desempeño, la capacidad de generalización y la robustez de diferentes algoritmos de clasificación de perturbaciones de calidad de energía. Se consideran tres dominios independientes de señales eléctricas: SEED-PQD, PQD-16c y un conjunto de señales sintéticas generado para el estudio. Se analizan ocho clases presentes en los tres dominios: Normal, Sag, Swell, Interruption, Harmonics, Flicker, Oscillatory Transient y Notch. Estas categorías permiten establecer una correspondencia directa entre los conjuntos de datos para evitar la comparación de perturbaciones compuestas.

El estudio compara tres algoritmos de aprendizaje automático: Support Vector Machine (SVM), Random Forest (RF) y XGBoost (XGB); y dos arquitecturas de aprendizaje profundo: una red neuronal convolucional unidimensional (1D-CNN) y una arquitectura híbrida CNN-LSTM. Se realiza evaluación dentro del mismo dominio (in-domain), transferencia directa entre dominios diferentes (cross-domain), análisis de suficiencia de datos, evaluación de robustez frente a ruido blanco gaussiano aditivo (AWGN) y análisis descriptivo del costo computacional.

Los experimentos se implementan en Python 3 mediante Google Colab, utilizando NumPy, pandas, SciPy, scikit-learn, XGBoost, Matplotlib y TensorFlow/Keras; para las arquitecturas profundas se utiliza TensorFlow 2.20.0. Las semillas aleatorias son controladas de acuerdo con cada experimento para favorecer la reproducibilidad. Debido a que Google Colab asigna recursos computacionales de forma dinámica y la configuración exacta de CPU, GPU y memoria RAM de las sesiones originales no fue almacenada de manera persistente, estas características no se recuperan con fiabilidad para el presente manuscrito. En consecuencia, los tiempos de entrenamiento e inferencia reportados se interpretan exclusivamente como indicadores descriptivos del entorno de ejecución utilizado y no como un benchmark absoluto de hardware.

Población, muestra y criterios de selección

La población de análisis está constituida por las señales disponibles en tres dominios: SEED-PQD-v1 [11], asociado al sistema XPQRS descrito en [12]; PQD-16c [13]; y un conjunto sintético generado en esta investigación. SEED-PQD-v1 contiene originalmente 17 clases, 1.000 señales por clase, 100 muestras por señal y una frecuencia de muestreo de 5 kHz [11], [12]. PQD-16c contiene 16 clases, 100 registros por clase, 640 muestras por registro y una frecuencia de muestreo de 3,2 kHz [13]. Después del proceso de armonización SEED-PQD aporta 8.000 señales, PQD-16c aporta 800 señales y el conjunto sintético 8.000 señales (ver Tabla 1). Cada registro está representado por un vector de 64

muestras y una etiqueta correspondiente a una de las ocho clases seleccionadas.

En SEED-PQD, las señales originalmente representadas mediante 100 muestras por ciclo a 5 kHz son remuestreadas a 64 muestras mediante remuestreo polifásico. En PQD-16c, cada registro contiene diez ciclos; para evitar considerar ciclos procedentes del mismo registro como observaciones independientes, se selecciona un único ciclo representativo mediante un criterio independiente de la etiqueta, basado en la desviación respecto a la componente fundamental y en cambios de amplitud. El conjunto sintético es generado directamente bajo la representación común de 64 muestras por ciclo.

Tabla I. Información de las bases de datos

Dominio	Señales	Señales por clase	Longitud	f_s	Clase
SEED-PQD	8.000	1.000	64	3.200 Hz	8
PQD-16c	800	100	64	3.200 Hz	8
Synthetic-PQD	8.000	1.000	64	3.200 Hz	8

Como criterios de inclusión se consideran señales pertenecientes a alguna de las ocho clases comunes, con correspondencia inequívoca entre dominios y sin valores faltantes o infinitos después del preprocesamiento. Se excluyeron las categorías no compartidas entre los tres conjuntos y las perturbaciones compuestas cuya equivalencia no podía garantizarse. La validación de integridad confirma ausencia de valores NaN e infinitos en los conjuntos finales.

En una primera etapa se construye un régimen controlado de 100 señales por clase y dominio, correspondiente a 800 registros por conjunto, con el propósito de comparar los modelos bajo una cantidad equivalente de observaciones. Luego se realiza el experimento principal utilizando todos los datos disponibles. Para este último, cada dominio se divide mediante muestreo estratificado en 80 % para entrenamiento y 20 % para prueba, con semilla 42. Así, SEED-PQD y Synthetic-PQD disponen de 6.400 señales de entrenamiento y 1.600 de prueba, mientras que PQD-16c dispone de 640 y 160 señales, respectivamente.

Durante la auditoría de duplicados se comparan las formas de onda completas dentro de cada clase. En los conjuntos controlados de SEED-PQD y PQD-16c, las 100 observaciones correspondientes a la clase Normal representan una única forma de onda repetida, es decir, se identifican 99 registros duplicados en cada dominio. En Synthetic-PQD, por el contrario, las 100 señales Normal son distintas. Dado que la partición estratificada 80/20 se efectúa a nivel de registro, en SEED-PQD y PQD-16c quedaron 80 registros Normal en entrenamiento y 20 en prueba, aunque todos correspondían a la misma forma de onda. Debido a que este solapamiento podía favorecer artificialmente el desempeño de la clase Normal, se realiza un análisis de sensibilidad complementario. Para este análisis no se reentrenan los clasificadores; se conservan los modelos obtenidos con las ocho clases y se recalculan las métricas después de retirar de la evaluación todas las observaciones cuya etiqueta verdadera correspondía a Normal. De esta manera, el

análisis permite estimar en qué medida la inclusión de esta categoría influía sobre las métricas agregadas, aunque no elimina el posible efecto de los registros duplicados durante el proceso de entrenamiento.

Variables, categorías e instrumentos

La variable de respuesta principal es el desempeño de clasificación, medido mediante accuracy, balanced accuracy, precisión macro, recall macro, Macro-F1 y coeficiente de correlación de Matthews (MCC). Macro-F1 se utiliza como métrica principal de comparación debido a que pondera de manera equivalente el desempeño de cada clase y permite realizar comparaciones entre dominios sin depender únicamente del número total de observaciones correctamente clasificadas. También se obtienen matrices de confusión y métricas específicas por clase.

Las variables experimentales son el algoritmo de clasificación, el dominio de entrenamiento, el dominio de evaluación, la cantidad de datos disponible para entrenamiento y la relación señal-ruido utilizada en las pruebas de robustez. La generalización se cuantifica mediante el desempeño cross-domain y la diferencia entre el Macro-F1 obtenido dentro del dominio y el alcanzado sobre un dominio externo:

$$Ggap = F1_{in-domain} - F1_{cross-domain}.$$

Para los algoritmos clásicos se extraen 13 características de cada señal: media, desviación estándar, varianza, valor RMS, valor pico, amplitud pico a pico, asimetría, curtosis, factor de cresta, energía, entropía espectral, magnitud de la componente fundamental de 50 Hz y distorsión armónica total (THD). La componente fundamental y el THD se calculan mediante la transformada rápida de Fourier; para el THD se consideran los armónicos tercero, quinto, séptimo y noveno cuando estaban contenidos dentro del rango espectral disponible.

SVM utiliza un escalamiento StandardScaler aprendido exclusivamente a partir de los datos de entrenamiento, seguido de un kernel RBF con $C=1$ y $\gamma=scale$. Random Forest se configura con 300 árboles. XGBoost emplea 300 estimadores, profundidad máxima de 6, tasa de aprendizaje de 0,05, subsample=0.8 y colsample_bytree=0.8.

La 1D-CNN recibe directamente las señales de dimensión (64\1). Su arquitectura está compuesta por tres capas convolucionales con 32, 64 y 128 filtros, tamaños de kernel de 5, 3 y 3, respectivamente; las dos primeras son seguidas por normalización por lotes y max pooling. Posteriormente se emplea global average pooling, una capa densa de 64 neuronas, dropout de 0,30 y una capa de salida softmax de ocho neuronas. La arquitectura tiene 40.264 parámetros totales.

La CNN-LSTM utiliza dos bloques convolucionales de 32 y 64 filtros, seguidos por

normalización por lotes y max pooling. La salida es procesada por una capa LSTM de 64 unidades, una capa densa de 64 neuronas, dropout de 0,30 y una capa softmax de ocho neuronas. Esta configuración presenta 44.488 parámetros totales.

Procedimiento y análisis de datos

Los modelos de aprendizaje automático son entrenados utilizando las 13 características previamente descritas. En el experimento principal, cada modelo es ajustado una única vez sobre el 80 % correspondiente al dominio fuente. Para la evaluación in-domain se utiliza el 20 % reservado como prueba. Para la evaluación cross-domain, el modelo entrenado sobre un dominio fuente es aplicado directamente al conjunto completo de cada dominio externo, sin utilizar observaciones del dominio objetivo durante el entrenamiento, selección de hiperparámetros o construcción del preprocesamiento.

Para los modelos profundos, el 80 % destinado al entrenamiento es subdividido nuevamente de manera estratificada, reservando 20 % de dicho subconjunto para validación y early stopping. Por tanto, aproximadamente 64 % del conjunto total es utilizado directamente para actualización de pesos, 16 % para validación y 20 % para prueba in-domain. La normalización se efectúa mediante la media y desviación estándar calculadas exclusivamente con el subconjunto de entrenamiento del dominio fuente; estos mismos parámetros se aplican posteriormente a validación, prueba y dominios externos. No se realiza normalización independiente por señal.

Las redes se entrenan con optimizador Adam y pérdida de entropía cruzada categórica dispersa, con máximo 150 épocas y tamaño de lote de 32. Se utiliza early stopping sobre la pérdida de validación con paciencia de 15 épocas y restauración de los mejores pesos. Asimismo, se emplea reducción adaptativa de la tasa de aprendizaje con factor 0,5, paciencia de siete épocas y tasa mínima de 10^{-6} . Los experimentos definitivos de 1D-CNN y CNN-LSTM se repiten con las semillas 42, 123 y 2026.

La sensibilidad a la disponibilidad de datos se examina específicamente para la 1D-CNN en los dominios SEED-PQD y Synthetic-PQD mediante un experimento independiente del protocolo principal. Para cada dominio se reserva una sola vez un conjunto de prueba fijo de 100 señales por clase, equivalente a 800 señales, utilizando muestreo estratificado con semilla 42. Este conjunto permanece invariable durante todas las condiciones del experimento y no participó en el entrenamiento ni en la selección de los subconjuntos.

A partir de las observaciones restantes se construyeron subconjuntos de entrenamiento de 80, 120, 160 y 200 señales por clase. Los subconjuntos no fueron anidados; para cada combinación de tamaño y semilla se realiza una nueva selección aleatoria estratificada sin reemplazo a partir del mismo pool disponible. Por tanto, las 80 señales empleadas en una condición no necesariamente están contenidas dentro de las 120, 160 o 200 señales utilizadas en las demás condiciones.

Cada tamaño de entrenamiento se evalúa con tres semillas independientes: 42, 123 y 2026. Estas semillas controlaron la selección del subconjunto de entrenamiento y la inicialización estocástica de la red. Dentro de cada repetición, el subconjunto seleccionado se divide nuevamente para entrenamiento y validación manteniendo el procedimiento de normalización, early stopping y optimización descrita previamente. El conjunto de prueba permanece idéntico entre tamaños y semillas, por lo que las diferencias observadas se atribuyen a la cantidad de datos disponibles para entrenamiento y a la variabilidad asociada a las distintas ejecuciones.

La robustez se evalúa introduciendo AWGN exclusivamente durante la fase de prueba. Se analizan relaciones señal-ruido de 30, 20, 15, 10, 5 y 0 dB, además de la condición limpia. Para cada señal, la potencia objetivo del ruido se obtiene mediante

$$P_n = \frac{P_x}{10^{SNR/10}} = P_x 10^{-SNR/10}$$

donde P_x corresponde a la potencia media de la señal. Para garantizar comparaciones pareadas, en cada repetición se genera una realización base de ruido gaussiano estándar que posteriormente fue escalada para obtener cada nivel de SNR. Se utilizan cinco semillas de ruido independientes: 101, 202, 303, 404 y 505.

En los modelos clásicos el AWGN se añade a la señal cruda antes de volver a extraer las 13 características. En las redes profundas, las señales contaminadas se normalizan utilizando la media y desviación estándar obtenidas del entrenamiento limpio. La degradación se cuantificó mediante

$$\Delta F1(SNR) = F1_{clean} - F1_{SNR},$$

y mediante la retención relativa:

$$R(SNR) = \frac{F1_{SNR}}{F1_{clean}}.$$

Adicionalmente, se calcula el área normalizada bajo la curva de Macro-F1 entre 0 y 30 dB como indicador global de robustez y el área bajo la curva de retención como medida de robustez relativa.

Para el análisis estadístico de robustez, la variable de respuesta es el Macro-F1 obtenido bajo cada condición de ruido. Se definen 18 bloques emparejados correspondientes a la combinación de los tres dominios de evaluación (SEED-PQD, PQD-16c y Synthetic-PQD), con seis niveles de relación señal-ruido (30, 20, 15, 10, 5 y 0 dB). En cada bloque se comparan los cinco clasificadores bajo la misma combinación dominio-SNR. La

condición limpia no se incluye en esta prueba, dado que el análisis está orientado al comportamiento bajo AWGN. Para los modelos clásicos, el Macro-F1 de cada condición corresponde al promedio de cinco realizaciones independientes de ruido, generadas con las semillas 101, 202, 303, 404 y 505. Para 1D-CNN y CNN-LSTM se combinan tres semillas de entrenamiento (42, 123 y 2026) con las cinco semillas de ruido, obteniéndose 15 evaluaciones por condición antes de calcular el valor medio utilizado en el análisis.

Las diferencias globales entre los cinco algoritmos se evalúan mediante la prueba no paramétrica de Friedman, con un nivel de significancia de $\alpha=0,05$. El tamaño de efecto global se estima mediante el coeficiente de concordancia Wde Kendall. Cuando la prueba global resultó significativa, se realizaron las diez comparaciones pareadas posibles mediante la prueba de rangos con signo de Wilcoxon, aplicando la corrección de Holm para controlar el error por comparaciones múltiples. Para cada contraste se reportan el valor de psin ajustar, el valor de pajustado mediante Holm y la correlación biserial por rangos r_{rb} como tamaño de efecto. La generalización cross-domain se analiza independientemente mediante Friedman sobre las seis combinaciones fuente-destino; al no obtenerse significancia global, no se realizan pruebas post-hoc para este experimento.

El costo computacional se caracteriza mediante tiempo de entrenamiento y tiempo de inferencia por muestra registrados con `time.perf_counter()`. Para las arquitecturas profundas también se cuantifica el número de parámetros, mientras que en SVM, Random Forest y XGBoost se utilizan como indicadores estructurales el número de vectores de soporte, el número total de nodos y el número de árboles, respectivamente.

Resultados

Desempeño de clasificación en condiciones in-domain

La evaluación inicial se realizó utilizando datos de prueba procedentes del mismo dominio empleado para el entrenamiento. La Tabla 2 presenta las métricas obtenidas por los cinco modelos en los tres conjuntos de datos. Para los modelos de aprendizaje profundo se reporta media $\pm$ desviación estándar de tres ejecuciones con diferentes semillas; los modelos de aprendizaje automático corresponden a la ejecución definida. En SEED, los clasificadores alcanzan valores elevados de Macro-F1, entre 0,9365 y 0,9883. CNN-LSTM presentó el mayor promedio, seguido por XGBoost, 1D-CNN y Random Forest. Un comportamiento similar se observa en el conjunto sintético, donde CNN-LSTM alcanzó un Macro-F1 de $0,9908 \pm 0,0052$, mientras que XGBoost, Random Forest y 1D-CNN superaron 0,97. En PQD16, Random Forest obtiene el mayor Macro-F1, con 0,8404, seguido por XGBoost con 0,8243 y SVM con 0,7424. Para 1D-CNN, las tres semillas produjeron Macro-F1 de 0,7754, 0,1933 y 0,8416; para CNN-LSTM se obtuvieron 0,8298, 0,1612 y 0,1994, con desviaciones estándar 0,3567 y 0,3755, respectivamente.

Tabla II. Desempeño in-domain.

Modelo	Dominio	Balanced Accuracy	Macro-F1	MCC
SVM	SEED	0,9363	0,9365	0,9282
Random Forest	SEED	0,9838	0,9837	0,9814
XGBoost	SEED	0,9875	0,9875	0,9857
1D-CNN	SEED	0,9842 $\pm$ 0,0019	0,9841 $\pm$ 0,0019	0,9819 $\pm$ 0,0022
CNN-LSTM	SEED	0,9883 $\pm$ 0,0022	0,9883 $\pm$ 0,0022	0,9867 $\pm$ 0,0025
SVM	PQD16	0,7563	0,7424	0,7299
Random Forest	PQD16	0,8438	0,8404	0,8224
XGBoost	PQD16	0,8250	0,8243	0,8012
1D-CNN	PQD16	0,6417 $\pm$ 0,2925	0,6034 $\pm$ 0,3567	0,6156 $\pm$ 0,2956
CNN-LSTM	PQD16	0,4375 $\pm$ 0,3509	0,3968 $\pm$ 0,3755	0,3851 $\pm$ 0,3862
SVM	Synthetic	0,9019	0,9021	0,8892
Random Forest	Synthetic	0,9775	0,9775	0,9743
XGBoost	Synthetic	0,9825	0,9825	0,9800
1D-CNN	Synthetic	0,9767 $\pm$ 0,0237	0,9768 $\pm$ 0,0235	0,9734 $\pm$ 0,0270
CNN-LSTM	Synthetic	0,9908 $\pm$ 0,0052	0,9908 $\pm$ 0,0052	0,9895 $\pm$ 0,0059

Generalización entre dominios

La Tabla III muestra las métricas obtenidas cuando cada modelo fue evaluado sobre un dominio diferente al del entrenamiento. Para las arquitecturas profundas corresponden al promedio de las tres semillas.

SVM presenta el mayor Macro-F1 cross-domain promedio, con 0,6407, seguido por 1D-CNN con 0,5813. Los valores promedio para Random Forest, XGBoost y CNN-LSTM fueron 0,4940, 0,4807 y 0,4760, respectivamente (ver Figura 1).

El comportamiento dependió de la dirección de transferencia. Cuando SEED-PQD fue utilizado como dominio fuente, SVM obtuvo un Macro-F1 de 0,7512 sobre Synthetic-PQD, mientras que CNN-LSTM alcanzó 0,6411 sobre PQD-16c. Cuando el entrenamiento se realizó con Synthetic-PQD, 1D-CNN obtuvo 0,8905 sobre SEED-PQD y 0,6167 sobre PQD-16c. Estas diferencias se interpretan de manera descriptiva, dado que la prueba global de Friedman no identificó diferencias estadísticamente significativas entre los clasificadores en las seis transferencias evaluadas.

Tabla III. Desempeño en las seis transferencias cross-domain.

Entrenamiento $\rightarrow$ prueba	SVM	Random Forest	XGBoost	1D-CNN	CNN-LSTM
PQD16 $\rightarrow$ SEED	0,6226 / 0,6115 / 0,5791	0,3993 / 0,3206 / 0,3261	0,3659 / 0,2916 / 0,2901	0,4862 $\pm$ 0,1740 / 0,4279 $\pm$ 0,2184 / 0,4436 $\pm$ 0,1678	0,3006 $\pm$ 0,1982 / 0,2605 $\pm$ 0,1797 / 0,2131 $\pm$ 0,2357
PQD16 $\rightarrow$ Synthetic	0,5446 / 0,5146 / 0,4895	0,4718 / 0,4579 / 0,4063	0,4146 / 0,4058 / 0,3404	0,3399 $\pm$ 0,0835 / 0,3098 $\pm$ 0,1449 / 0,2703 $\pm$ 0,0692	0,2003 $\pm$ 0,0762 / 0,1629 $\pm$ 0,0749 / 0,0946 $\pm$ 0,0946
SEED $\rightarrow$ PQD16	0,6638 / 0,5935 / 0,6339	0,4875 / 0,3768 / 0,4731	0,4888 / 0,3562 / 0,4572	0,7054 $\pm$ 0,0204 / 0,6259 $\pm$ 0,0186 / 0,6844 $\pm$ 0,0232	0,7246 $\pm$ 0,0081 / 0,6411 $\pm$ 0,0065 / 0,7047 $\pm$ 0,0104

SEED → Synthetic	0,7585 / 0,7512 / 0,7293	0,7296 / 0,6846 / 0,7008	0,7493 / 0,7338 / 0,7211	0,6503 ± 0,0025 / 0,6172 ± 0,0056 / 0,6272 ± 0,0024	0,5025 ± 0,0238 / 0,4488 ± 0,0287 / 0,4671 ± 0,0216
Synthetic → PQD16	0,5800 / 0,5437 / 0,5274	0,4575 / 0,3726 / 0,3949	0,4150 / 0,3264 / 0,3445	0,6883 ± 0,0120 / 0,6167 ± 0,0118 / 0,6588 ± 0,0139	0,6267 ± 0,0102 / 0,5560 ± 0,0094 / 0,5867 ± 0,0125
Synthetic → SEED	0,8373 / 0,8295 / 0,8157	0,7820 / 0,7513 / 0,7590	0,7913 / 0,7705 / 0,7676	0,8907 ± 0,0130 / 0,8905 ± 0,0122 / 0,8766 ± 0,0145	0,7985 ± 0,0415 / 0,7865 ± 0,0384 / 0,7761 ± 0,0484
Promedio	0,6678 / 0,6407 / 0,6291	0,5546 / 0,4940 / 0,5100	0,5375 / 0,4807 / 0,4868	0,6268 / 0,5813 / 0,5935	0,5255 / 0,4760 / 0,4737

La reducción del desempeño respecto a la evaluación *in-domain* se cuantificó mediante el *generalization gap*. El promedio de esta diferencia fue de 0,2197 para SVM, 0,2735 para 1D-CNN, 0,3160 para CNN-LSTM, 0,4399 para Random Forest y 0,4507 para XGBoost. La transferencia con menor brecha en 1D-CNN fue Synthetic-SEED:

$$G_{gap} = 0,0863 \pm 0,0170,$$

con:

$$Macro-F1_{cross-domain} = 0,8905 \pm 0,0122.$$

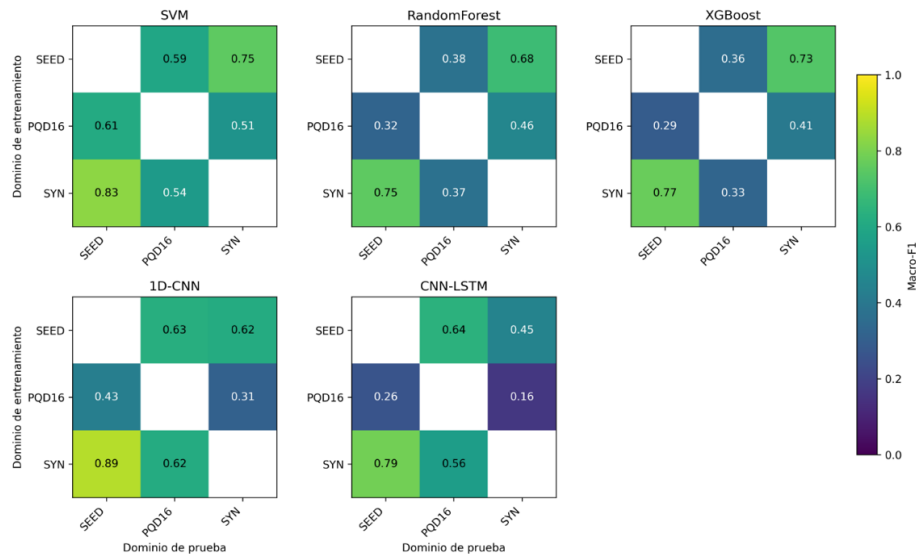


Figura 1. Generalización cross-domain de los cinco clasificadores entre los dominios SEED, PQD16 y Synthetic.

En el análisis por clase de los modelos clásicos, *Flicker* presentó el menor Macro-F1 promedio *cross-domain* (0,2864), seguido por *Oscillatory Transient* (0,4507). Las clases con mayor desempeño promedio fueron *Swell* (0,7321), *Normal* (0,7065) y *Harmonics* (0,6923).

La comparación global de los cinco clasificadores mediante la prueba de Friedman, utilizando como bloques las seis combinaciones fuente-destino, produjo:

$$\chi^2_2(4)=7,2000, \\ p=0,1257$$

y un coeficiente de concordancia de Kendall

$$W = 0,3000.$$

Por tanto, las diferencias observadas entre los valores de Macro-F1 cross-domain son descriptivas y no alcanzan significancia estadística global al nivel $\alpha=0,05$. En consecuencia, no se realizan comparaciones post-hoc de Wilcoxon para este experimento. Los rangos promedio fueron 1,8333 para SVM, 2,3333 para 1D-CNN, 3,5000 para Random Forest, 3,5000 para CNN-LSTM y 3,8333 para XGBoost.

El análisis de sensibilidad mostró cambios reducidos en el Macro-F1 promedio de las seis transferencias cross-domain al excluir la clase Normal de la evaluación (ver Tabla 4). En SVM, el valor pasó de 0,5998 a 0,5926 ($\Delta=-0,0072$); en Random Forest, de 0,6082 a 0,5742 ($\Delta=-0,0340$); y en XGBoost, de 0,5362 a 0,5373 ($\Delta=+0,0011$). Aunque Random Forest presentó la mayor reducción, la exclusión de la categoría Normal no produjo cambios sustanciales en la tendencia general observada entre los tres clasificadores clásicos. Estos resultados indican que las métricas cross-domain no dependen exclusivamente del desempeño sobre la clase duplicada. No obstante, debido a que los modelos no fueron reentrenados después de eliminar los duplicados, este análisis no permite descartar completamente su posible influencia sobre el aprendizaje ni cuantificar con precisión la magnitud de una eventual sobreestimación del desempeño in-domain.

Tabla IV. Análisis de sensibilidad frente a la exclusión de la clase Normal.

Modelo	Macro-F1 con Normal	Macro-F1 sin Normal	Cambio
SVM	0,5998	0,5926	-0,0072
Random Forest	0,6082	0,5742	-0,0340
XGBoost	0,5362	0,5373	+0,0011

Influencia de la cantidad de datos sobre la 1D-CNN

La disponibilidad de datos muestra un efecto importante sobre el entrenamiento de la 1D-CNN. Para estudiar este comportamiento se emplean SEED-PQD y Synthetic-PQD. En cada dominio se mantiene fijo el mismo conjunto de prueba de 100 señales por clase para todas las condiciones, mientras que el conjunto de entrenamiento se construye independientemente para 80, 120, 160 y 200 señales por clase y para las tres semillas evaluadas (ver Tabla V).

Tabla V. Macro-F1 de 1D-CNN en función de la cantidad de datos disponibles por clase.

Señales por clase	SEED	Synthetic
80	0,1638 $\pm$ 0,0543	0,2153 $\pm$ 0,0638
120	0,6826 $\pm$ 0,4882	0,4347 $\pm$ 0,4275
160	0,9711 $\pm$ 0,0123	0,9428 $\pm$ 0,0055

200	$0,9710 \pm 0,0052$	$0,9506 \pm 0,0051$
-----	---------------------	---------------------

Con 80 señales disponibles por clase, el Macro-F1 promedio fue inferior a 0,22 en ambos dominios. Al aumentar a 120 señales se incrementó el desempeño medio, pero también se observó una elevada variabilidad entre ejecuciones, con desviaciones estándar de 0,4882 en SEED y 0,4275 en Synthetic.

A partir de 160 señales disponibles por clase, la variabilidad disminuyó de forma pronunciada. SEED-PQD alcanzó un Macro-F1 de $0,9711 \pm 0,0123$ y Synthetic-PQD de $0,9428 \pm 0,0055$. Con 200 señales por clase, los valores fueron $0,9710 \pm 0,0052$ y $0,9506 \pm 0,0051$, respectivamente.

En el análisis exploratorio con volúmenes mayores, la 1D-CNN alcanzó 0,9824 con 400 señales por clase y 0,9875 con 800 en SEED; en Synthetic los valores correspondientes fueron 0,9649 y 0,9888 (ver Figura 2).

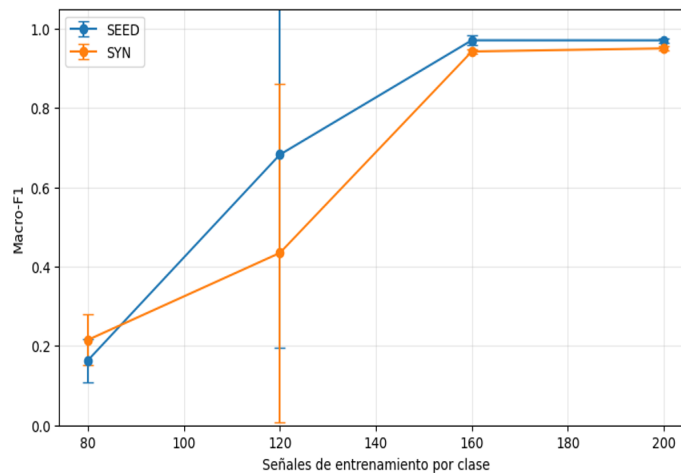


Figura 2. Sensibilidad de la 1D-CNN al tamaño del conjunto de entrenamiento.

Robustez frente a ruido AWGN

Las figuras 3 a 5 presentan la evolución de Macro-F1 a medida que disminuyó la relación señal-ruido. La Tabla 6 resume el desempeño medio sobre los tres dominios para cada nivel de SNR.

SVM alcanzó el mayor Macro-F1 promedio a 30, 20 y 15 dB. A 30 dB obtuvo 0,8449, frente a 0,6509 de XGBoost y 0,6424 de Random Forest. A 20 dB la diferencia se mantuvo, con 0,6039 para SVM, 0,4324 para Random Forest y 0,4318 para XGBoost.

La ventaja de SVM disminuyó con la reducción de SNR. A 10 dB Random Forest obtuvo el mayor valor promedio, 0,2893, frente a 0,2850 para SVM. En condiciones extremas de 0 dB, XGBoost presentó el mayor Macro-F1 promedio, con 0,1376.

Las arquitecturas profundas mostraron una disminución pronunciada respecto a la condición limpia. En SEED-PQD, 1D-CNN pasó de un Macro-F1 de 0,9841 en condición limpia a $0,6393 \pm 0,0099$ a 30 dB y $0,3267 \pm 0,0439$ a 20 dB. CNN-LSTM pasó de 0,9883 en condición limpia a $0,6662 \pm 0,0180$ a 30 dB y $0,3862 \pm 0,0747$ a 20 dB. En Synthetic-PQD se observó una degradación similar. A 20 dB, 1D-CNN obtuvo un Macro-F1 de $0,2500 \pm 0,0579$, mientras que CNN-LSTM alcanzó $0,3575 \pm 0,0815$.

Para sintetizar el comportamiento en el intervalo completo de 0–30 dB se calcularon el área bajo la curva de robustez y el área de retención del desempeño (ver Tabla 7).

La comparación de las 18 condiciones formadas por tres dominios y seis niveles de SNR mediante la prueba de Friedman produjo:

$$\chi^2_2 = 37,9111,$$

$$p = 1,2 \times 10^{-7},$$

y:

$$W = 0,5265.$$

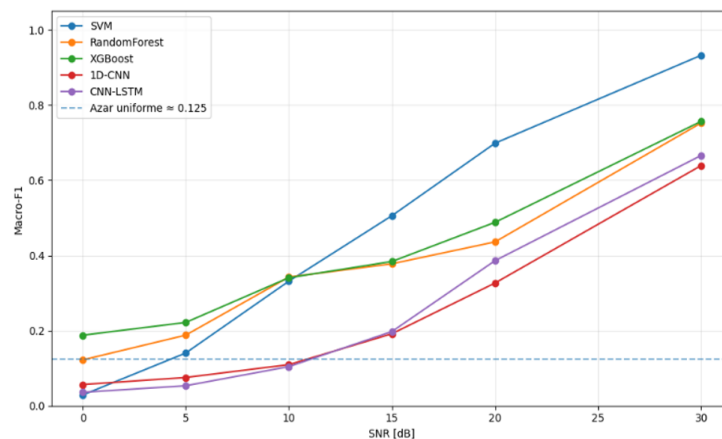


Figura 3. Robustez frente a AWGN - SEDD

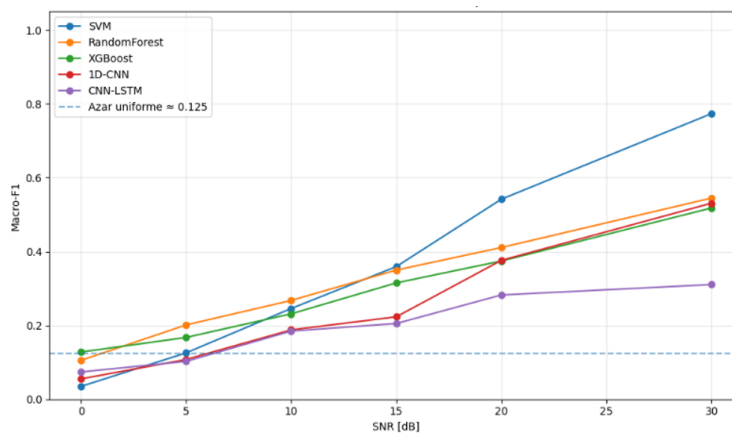


Figura 4. Robustez frente a AWGN - PQD16

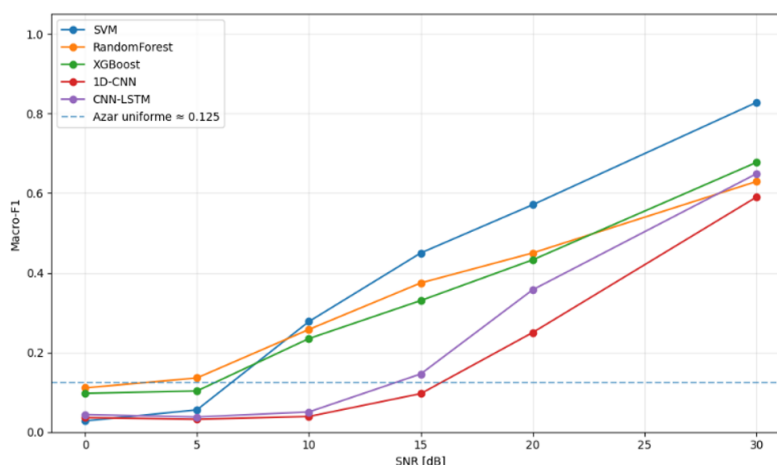


Figura 5. Robustez frente a AWGN - SYN

Tabla VI. Macro-F1 promedio sobre los tres dominios bajo diferentes niveles de AWGN.

Modelo	30 dB	20 dB	15 dB	10 dB	5 dB	0 dB
SVM	0,8449	0,6039	0,4383	0,2850	0,1072	0,0305
Random Forest	0,6424	0,4324	0,3675	0,2893	0,1750	0,1127
XGBoost	0,6509	0,4318	0,3431	0,2690	0,1641	0,1376
1D-CNN	0,5870	0,3177	0,1705	0,1121	0,0717	0,0493
CNN-LSTM	0,5420	0,3421	0,1829	0,1131	0,0645	0,0512

Tabla VII. Indicadores globales de robustez.

Modelo	Macro-F1 medio bajo ruido	AUC robustez	AUC retención
SVM	0,3850	0,4328	0,5039
Random Forest	0,3366	0,3632	0,3895
XGBoost	0,3328	0,3573	0,3831
1D-CNN	0,2181	0,2404	0,3033
CNN-LSTM	0,2160	0,2402	0,3488

La comparación estadística se realiza sobre los 18 bloques emparejados definidos por las tres bases de datos y los seis niveles de SNR. La variable analizada es el Macro-F1 medio de cada clasificador en cada combinación dominio-SNR. La prueba de Friedman muestra diferencias globales estadísticamente significativas entre los cinco clasificadores, $\chi^2_2(4)=37,9111$, $p=1,17 \times 10^{-7}$, con un coeficiente de concordancia de Kendall $W=0,5265$.

Los rangos promedio fueron 2,0000 para Random Forest, 2,1667 para SVM, 2,3333 para XGBoost, 4,2222 para CNN-LSTM y 4,2778 para 1D-CNN. Dado que en la prueba de Friedman los rangos menores corresponden a un mejor posicionamiento relativo dentro de cada bloque, los tres modelos clásicos ocuparon, en promedio, posiciones más favorables bajo las condiciones de ruido evaluadas. Estos rangos representan el orden relativo de los clasificadores dentro de las 18 condiciones y, por tanto, no son equivalentes a los indicadores de AUC de la Tabla 6, los cuales incorporan la magnitud del Macro-F1 a lo largo de todo el intervalo de SNR.

Debido a la significancia de la prueba global, se realizan comparaciones post-hoc mediante Wilcoxon con corrección de Holm. No se encontraron diferencias estadísticamente significativas entre SVM, Random Forest y XGBoost, ni entre 1D-CNN y CNN-LSTM. En contraste, todas las comparaciones entre los clasificadores clásicos y las dos arquitecturas profundas fueron significativas después del ajuste por comparaciones múltiples. Los resultados completos, incluyendo valores de p ajustados y tamaños de efecto, se presentan en la Tabla VIII.

Costo computacional

La Tabla 9 resume los tiempos registrados durante los experimentos. Estos valores corresponden al entorno de ejecución utilizado y se presentan como indicadores descriptivos.

Entre los algoritmos clásicos, SVM presentó el menor tiempo medio de entrenamiento, mientras XGBoost registró el menor tiempo de inferencia por muestra. Las arquitecturas profundas requirieron tiempos de entrenamiento superiores; 1D-CNN promedió 98,32 s y CNN-LSTM 159,23 s.

Tabla VIII. Comparaciones post-hoc de Wilcoxon para robustez frente a AWGN.

Comparación	W	p	p-Holm	r_{rb}	Significativa
SVM – Random Forest	55	0,196388	0,589165	0,3567	No
SVM – XGBoost	48	0,108383	0,433533	0,4386	No
SVM – 1D-CNN	9	0,000252	0,001259	0,8947	Sí
SVM – CNN-LSTM	8	0,000191	0,001144	0,9064	Sí
Random Forest – XGBoost	72	0,579842	0,990158	0,1579	No
Random Forest – 1D-CNN	0	0,000008	0,000076	1,0000	Sí
Random Forest – CNN-LSTM	1	0,000015	0,000122	0,9883	Sí
XGBoost – 1D-CNN	3	0,000038	0,000267	0,9649	Sí
XGBoost – CNN-LSTM	0	0,000008	0,000076	1,0000	Sí
1D-CNN – CNN-LSTM	69	0,495079	0,990158	-0,1930	No

La 1D-CNN estuvo conformada por 40.264 parámetros, mientras que CNN-LSTM presentó 44.488 parámetros. Como medidas estructurales específicas de los algoritmos clásicos, el modelo SVM entrenado sobre SEED utilizó 1.928 vectores de soporte, Random Forest acumuló 91.282 nodos en sus árboles y XGBoost generó 2.400 árboles considerando el esquema multiclase.

Tabla IX. Tiempo de entrenamiento e inferencia registrado.

Modelo	Tiempo de entrenamiento (s)	Inferencia (ms/muestra)
SVM	0,3216	0,1561
Random Forest	3,2608	0,1878
XGBoost	3,6900	0,0416
1D-CNN	98,3223 ± 84,0734	0,5508 ± 0,4938
CNN-LSTM	159,2258 ± 124,7004	0,4414 ± 0,1988

Discusión

Los resultados muestran que alto desempeño in-domain no garantiza generalización ni robustez. XGBoost, 1D-CNN y CNN-LSTM alcanzaron Macro-F1 cercanos o superiores a 0,98 en SEED y Synthetic, pero su rendimiento disminuyó al cambiar de dominio o incorporar ruido AWGN. Este comportamiento se relaciona con los problemas de sobreajuste, dependencia de los datos y variaciones de distribución reportados en sistemas de clasificación [14]-[29].

El tamaño del conjunto de entrenamiento afectó a los modelos. La 1D-CNN presentó bajo desempeño y alta variabilidad con 80–120 señales por clase, mientras que a partir de 160 señales superó un Macro-F1 de 0,94 en SEED y Synthetic. En PQD16, con menos datos disponibles, las redes profundas mostraron mayor variabilidad entre semillas que los métodos clásicos. Estos resultados evidencian una sensibilidad de las arquitecturas profundas a la cantidad de datos disponible para entrenamiento. Aunque la representatividad de las muestras podría contribuir también a la estabilidad del modelo, este factor no fue cuantificado directamente en el presente estudio.

En las pruebas cross-domain, SVM registró el mayor Macro-F1 promedio (0,6407); sin embargo, la prueba de Friedman no mostró diferencias globales estadísticamente significativas entre los clasificadores. El desempeño también varió según la combinación origen-destino: la 1D-CNN entrenada con Synthetic-PQD alcanzó un Macro-F1 de 0,8905 sobre SEED-PQD, mientras que las transferencias hacia PQD-16c presentaron valores inferiores. Este patrón es consistente con la hipótesis de que las diferencias entre las distribuciones fuente y destino pueden afectar la capacidad de transferencia; no obstante, la similitud entre distribuciones no fue cuantificada directamente en este estudio, por lo que esta explicación debe interpretarse como una hipótesis derivada de los resultados y no como una relación demostrada.

Las diferencias más claras entre los grupos de modelos aparecieron bajo AWGN. SVM registró los valores más altos de AUC de robustez (0,4328) y AUC de retención (0,5039). La prueba de Friedman mostró diferencias globales significativas y las comparaciones post-hoc identificaron diferencias significativas entre los modelos clásicos y las arquitecturas profundas, mientras que no se detectaron diferencias significativas entre SVM, Random Forest y XGBoost. Dado que las redes fueron entrenadas únicamente con señales limpias, estos resultados caracterizan su respuesta frente a una degradación no observada durante el entrenamiento. En términos descriptivos, SVM reunió el mayor Macro-F1 cross-domain promedio, los mayores valores de AUC de robustez y retención y el menor tiempo de entrenamiento entre los modelos clásicos. Sin embargo, el estudio no definió un índice multicriterio que combine formalmente generalización, robustez y costo computacional; por tanto, estos resultados no permiten establecer un balance global superior de un clasificador sobre los demás.

Limitaciones

Este estudio presenta limitaciones en la armonización de las señales a 64 muestras y solo se analizaron ocho clases comunes a los tres dominios y se detectaron formas de onda duplicadas en la clase Normal de SEED-PQD y PQD-16c. El análisis de sensibilidad excluyó esta categoría durante el recálculo de las métricas, pero los clasificadores no fueron reentrenados después de eliminar los registros duplicados; por tanto, no puede descartarse completamente su posible influencia sobre los resultados in-domain. La diferencia de tamaño entre SEED, Synthetic y PQD16 afectó principalmente la estabilidad de los modelos profundos. La generalización se evaluó con solo seis transferencias cross-domain. La robustez se estudió con ruido AWGN entre 0 y 30 dB, sin considerar otras degradaciones reales. Los tiempos computacionales obtenidos en Google Colab deben interpretarse como referencias descriptivas.

Conclusiones

La evaluación del desempeño in-domain, la generalización cross-domain y la robustez frente al ruido mostró que buen rendimiento en condiciones controladas no garantiza el mismo comportamiento cuando cambia el dominio de los datos o disminuye la calidad de la señal. Los modelos profundos alcanzaron valores de Macro-F1 superiores a 0,98 en SEED y Synthetic cuando tenían suficientes datos de entrenamiento. Sin embargo, su desempeño disminuyó al reducir las muestras e incorporar ruido AWGN. El análisis de suficiencia de datos mostró entrenamientos inestables con 80–120 señales por clase, mientras que a partir de 160 señales por clase los resultados fueron más estables y superaron un Macro-F1 de 0,94 en la configuración evaluada.

En las pruebas de generalización entre dominios, SVM obtuvo el mayor Macro-F1 promedio (0,6407) y presentó menores pérdidas de desempeño. Sin embargo, la prueba de Friedman no encontró diferencias globales estadísticamente significativas entre los cinco clasificadores en las seis transferencias cross-domain ($p=0,1257$). La 1D-CNN entrenada con Synthetic-PQD y evaluada sobre SEED-PQD alcanzó un Macro-F1 de $0,8905 \pm 0,0122$ y un generalization gap de $0,0863 \pm 0,0170$. Este resultado muestra que, en la combinación específica Synthetic-PQD-SEED-PQD, fue posible conservar una proporción elevada del desempeño in-domain. La posible influencia de la similitud entre ambos dominios constituye una interpretación plausible, pero no fue evaluada cuantitativamente en este estudio.

En la prueba de robustez frente a AWGN, SVM obtuvo los mejores indicadores globales, con un AUC de robustez de 0,4328 y un AUC de retención de 0,5039, y mantuvo un mejor desempeño bajo niveles de ruido leve y moderado. La prueba de Friedman confirmó diferencias significativas entre los clasificadores ($p<0,001$, $(W=0,5265)$). El análisis post-hoc mostró diferencias entre los modelos clásicos y las arquitecturas profundas, pero no entre SVM, Random Forest y XGBoost.

Los resultados no permiten identificar un clasificador superior en todos los escenarios evaluados. Las arquitecturas profundas alcanzaron valores de Macro-F1 cercanos o superiores a 0,98 en SEED-PQD y Synthetic-PQD cuando dispusieron de suficientes datos, pero mostraron una mayor degradación frente a AWGN. SVM registró el mayor Macro-F1 cross-domain promedio y los valores más altos de AUC de robustez y retención, aunque la superioridad cross-domain no fue estadísticamente significativa y no se definió un criterio multicriterio para integrar formalmente generalización, robustez y costo computacional. En consecuencia, la selección de un clasificador para perturbaciones de calidad de energía debe considerar conjuntamente el escenario de aplicación, la disponibilidad de datos, la transferencia entre dominios, la tolerancia al ruido y los requerimientos computacionales.

Disponibilidad de datos

Los datos utilizados en este estudio proceden de conjuntos de datos públicos previamente disponibles, incluyendo SEED-PQD [11] y PQD-16c [13], así como de un conjunto de señales sintéticas generado para esta investigación. Los resultados en el estudio están disponibles previa solicitud al autor de correspondencia.

Consideraciones éticas

El estudio utilizó conjuntos de datos secundarios de señales eléctricas y datos sintéticos, sin participación de seres humanos o animales ni tratamiento de información personal. Por esta razón, no requirió aprobación de un comité de ética ni consentimiento informado.

Referencias

- [1] A. A. Memon, M. A. Koondhar, S. F. Al-Gahtani, Z. M. S. Elbarbary, and Z. M. Alaas, "Comprehensive review of power quality disturbance detection and classification techniques," *Computers & Electrical Engineering*, vol. 126, Art. no. 110512, 2025, doi: 10.1016/j.compeleceng.2025.110512.
- [2] J. E. Caicedo, D. Agudelo-Martínez, E. Rivas-Trujillo, and J. Meyer, "A systematic review of real-time detection and classification of power quality disturbances," *Protection and Control of Modern Power Systems*, vol. 8, Art. no. 3, 2023, doi: 10.1186/s41601-023-00277-y.
- [3] J. C. Palomares-Salas, S. Aguado-González, and J. M. Sierra-Fernández, "Robustness of machine learning and deep learning models for power quality disturbance classification: A cross-platform analysis," *Applied Sciences*, vol. 15, no. 19, Art. no. 10602, 2025, doi: 10.3390/app151910602.
- [4] M. A. A. Baig, N. I. Ratyal, A. Amin, U. Jamil, S. Liaquat, H. M. Khalid, and M. F. Zia, "An ensemble deep CNN approach for power quality disturbance classification: A

technological route towards smart cities using image-based transfer," *Future Internet*, vol. 16, no. 12, Art. no. 436, 2024, doi: 10.3390/fi16120436.

- [5] M. Mosayebi, S. Azad, and M. T. Ameli, "Unknown power quality disturbances classification based on transfer learning approach with imbalanced data considerations," *Results in Engineering*, vol. 27, Art. no. 105865, 2025, doi: 10.1016/j.rineng.2025.105865.
- [6] U. Sipai, R. Jadeja, N. Kothari, T. Trivedi, and K. K. Ram, "Deep transfer learning approach for the classification of single and multiple power quality disturbances," *Scientific Reports*, vol. 15, Art. no. 34583, 2025, doi: 10.1038/s41598-025-18064-0.
- [7] I. Kapuza, E. Ginzburg-Ganz, R. Machlev, and Y. Levron, "Improving robustness of Transformers for power quality disturbance classification via optimized relevance maps," *Engineering Applications of Artificial Intelligence*, vol. 161, Art. no. 112138, 2025, doi: 10.1016/j.engappai.2025.112138.
- [8] X. Zhang, C. Jiang, M. Yu, X. Wen, J. Zhang, N. Rong, and S. Han, "Adversarial black-box attack and defense for convolutional neural network-based power quality disturbance classification," *Engineering Applications of Artificial Intelligence*, vol. 162, Art. no. 112411, 2025, doi: 10.1016/j.engappai.2025.112411.
- [9] E. García Rodríguez, E. Reyes Archundia, J. A. Gutiérrez Gnechi, O. I. Coronado Reyes, J. C. Olivares Rojas, and A. Méndez Patiño, "Detection and extraction of optimal features from power quality disturbances based on wavelet coefficient reconstruction and noise-robust methods," *Computers & Electrical Engineering*, vol. 127, Art. no. 110615, 2025, doi: 10.1016/j.compeleceng.2025.110615.
- [10] Q. Xu, F. Zhu, W. Jiang, X. Pan, P. Li, X. Zhou, and Y. Wang, "Efficient identification method for power quality disturbance: A hybrid data-driven strategy," *Processes*, vol. 12, no. 7, Art. no. 1395, 2024, doi: 10.3390/pr12071395.
- [11] M. U. Khan, S. Aziz, and A. Usman, "SEED_PQD_v1 (SEED—Power Quality Disturbance Dataset ver1)," *Zenodo, dataset*, 2024. [Online]. <https://zenodo.org/records/11843312?utm>
- [12] M. U. Khan, S. Aziz, and A. Usman, "XPQRS: Expert power quality recognition system for sensitive load applications," *Measurement*, vol. 216, Art. no. 112889, 2023, doi: 10.1016/j.measurement.2023.112889.
- [13] jkexin, "PQD-16c-Data: Power quality disturbance dataset," *Kaggle, dataset*. [Online]. <https://www.kaggle.com/datasets/jkexin/pqd-16c-data?utm>
- [14] D. Li, I. A. Channa, X. Chen, L. Dou, S. Khokhar, and N. Ab Azar, "A new deep learning

method for classification of power quality disturbances using DWT-MRA in utility smart grid," *Computers & Electrical Engineering*, vol. 117, Art. no. 109290, 2024, doi: 10.1016/j.compeleceng.2024.109290.

- [15] L. Chen, S. Chen, J. Xu, and C. Zhou, "Power quality disturbances identification based on deep neural network model of time-frequency feature fusion," *Electric Power Systems Research*, vol. 231, Art. no. 110283, 2024, doi: 10.1016/j.epsr.2024.110283.
- [16] T. Liao, W. Wang, and Y. Xing, "A method for disturbance identification in power quality based on cross-attention fusion of temporal and spatial features," *Electric Power Systems Research*, vol. 234, Art. no. 110560, 2024, doi: 10.1016/j.epsr.2024.110560.
- [17] P. Savaridass M and S. Esakkirajan, "Hilbert transform based combined 1D and 2D deep learning framework for power quality disturbance classification," *Electric Power Systems Research*, vol. 249, Art. no. 112031, 2025, doi: 10.1016/j.epsr.2025.112031.
- [18] M. A. A. Baig, N. I. Ratyal, A. Amin, U. Jamil, H. M. Khalid, and M. F. Zia, "A multi-modal deep learning framework for power quality disturbance classification: An integration of 1D time-series signals and 2D scalograms," *Computers & Electrical Engineering*, vol. 128, Art. no. 110716, 2025, doi: 10.1016/j.compeleceng.2025.110716.
- [19] P. Khetarpal, N. Nagpal, M. S. Al-Numay, P. Siano, Y. Arya, and N. Kassarwani, "Power quality disturbances detection and classification based on deep convolution auto-encoder networks," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3274732.
- [20] M. Liu, Y. Chen, Z. Zhang, and S. Deng, "Classification of power quality disturbance using segmented and modified S-transform and DCNN-MSVM hybrid model," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2022.3233767.
- [21] M. Abubakar, A. A. Nagra, M. Mudassar, M. Faheem, and M. Sohail, "High-precision identification of power quality disturbances based on discrete orthogonal S-transforms and compressed neural network methods," *IEEE Access*, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3304375.
- [22] V. Veeramsetty, A. Dhanush, A. Nagapradhyullatha, G. Rama Krishna, and S. R. Salkuti, "Power quality disturbances classification using autoencoder and radial basis function neural network," *International Journal of Emerging Electric Power Systems*, vol. 25, no. 6, pp. 817–842, 2024, doi: 10.1515/ijeeps-2023-0143.
- [23] S. Janthong and P. Phukpattaranont, "Identification of power quality disturbances in electrical distribution system using fast Fourier transforms and super learner ensembles," *Journal of Advanced Research in Applied Mechanics*, vol. 124, no. 1, pp. 39–60, 2024, doi: 10.37934/aram.124.1.3960.

- [24] H. Zhang, W. Wu, K. Li, X. Zheng, X. Xu, X. Wei, and C. Zhao, "Multi-strategy active learning for power quality disturbance identification," *Applied Soft Computing*, vol. 154, Art. no. 111326, 2024, doi: 10.1016/j.asoc.2024.111326.
- [25] A. Zlatkova and D. Taskovski, "A novel CNN-based framework for detection and classification of power quality disturbances: Exploring multi-class versus multi-label classification," *IEEE Access*, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3546520.
- [26] M. Alsabaan, A. Elsayed, A. Bondok, M. M. Badr, M. Mahmoud, T. Alshawhi, and M. I. Ibrahim, "Robust federated-learning-based classifier for smart grid power quality disturbances," *Sensors*, vol. 25, no. 22, Art. no. 6880, 2025, doi: 10.3390/s25226880.
- [27] H. Castro-Gutiérrez, C. Robles-Algarín, and L. L. Camargo-Ariza, "Análisis comparativo de los algoritmos de árbol de decisión y máquinas de vectores de soporte en la clasificación de impulsores de bomba," *Respuestas*, vol. 30, no. 1, pp. 95–106, 2025, doi: 10.22463/0122820X.5233.
- [28] D. C. Candia-Herrera, K. Y. Sánchez-Mojica, S. L. Zarama-Ortiz, and C. L. Díaz-Mesa, "Hacia un modelo universal de IA para la detección simultánea de COVID-19, neumonía bacteriana y tuberculosis en imágenes de rayos X: Desafíos y oportunidades," *Respuestas*, vol. 31, no. 1, pp. 47–61, 2026, doi: 10.22463/0122820X.5433.
- [29] M. Á. Flórez-Zambrano, F. A. León-García, J. L. Díaz-Rodríguez, and O. J. Suárez-Sierra, "Sistema en tiempo real para el reconocimiento de vocales en LSC mediante CNN y visión por computador," *Respuestas*, vol. 31, no. 2, 2026, doi: 10.22463/0122820X.5918.