



Pronóstico de Volatilidad del Bitcoin mediante Redes Neuronales Profundas: Una Comparación entre Arquitecturas Secuenciales y Métodos de Gradient Boosting

Juan Pablo Mazo Quintero
Danilo Rodriguez Arango

Trabajo de grado para optar al título de especialista en analítica y ciencia de datos

Asesor(a)
Gabriel Darío Uribe Guerra, PhD

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín
2026

Cita	(Mazo Quintero & Rodriguez Arango, 2026)
Referencia	Mazo Quintero, J. P., & Rodriguez Arango, D. (2026). <i>Pronóstico de Volatilidad del Bitcoin mediante Redes Neuronales Profundas: Una Comparación entre Arquitecturas Secuenciales y Métodos de Gradient Boosting</i> [Trabajo de grado]. Universidad de Antioquia, Medellín.
Estilo APA 7	



Facultad de Ingeniería.

Especialización en Analítica y Ciencia de Datos.



Repositorio Institucional: <http://bibliotecadigital.udea.edu.co>

Universidad de Antioquia – <https://www.udea.edu.co>

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Índice

Resumen	7
Abstract	7
1. Introducción	8
2. Metodología	12
2.1. Descripción de los datos	12
2.2. Preprocesamiento y limpieza de los datos	14
2.3. Método de análisis y modelado	15
2.4. Evaluación del modelo	15
3. Resultados y Discusión	16
4. Conclusiones	19
5. Referencias	21

Lista de tablas

1.	Partición cronológica de los datos.	14
2.	Desempeño fuera de muestra de los modelos evaluados.	16

Lista de figuras

1.	Precio de cierre, log returns y volatilidad realizada logarítmica.	13
2.	Predicciones en las últimas 500 horas del test set.	17

Siglas, acrónimos y abreviaturas

APA	American Psychological Association
API	Application Programming Interface
ARCH	Autoregressive Conditional Heteroskedasticity
ARIMA	AutoRegressive Integrated Moving Average
BTC	Bitcoin
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GRU	Gated Recurrent Unit
HAR	Heterogeneous Autoregressive
LightGBM	Light Gradient Boosting Machine
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MIDAS	Mixed Data Sampling
PhD	Philosophiae Doctor
QLIKE	Quasi-Likelihood
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RV	Realized Volatility
SHAP	SHapley Additive exPlanations
UdeA	Universidad de Antioquia
VaR	Value at Risk
XGBoost	Extreme Gradient Boosting

Pronóstico de Volatilidad del Bitcoin mediante Redes Neuronales Profundas: Una Comparación entre Arquitecturas Secuenciales y Métodos de Gradient Boosting

Juan Pablo Mazo Quintero

Danilo Rodríguez Arango

Resumen

La volatilidad del Bitcoin plantea uno de los retos más persistentes en la gestión de riesgo de los mercados de criptoactivos: su carácter discontinuo, su sensibilidad extrema a noticias y su estructura de dependencias no lineales hacen que ningún modelo capture de forma definitiva su comportamiento. Este trabajo examina la capacidad predictiva de distintas aproximaciones econométricas, de aprendizaje automático y de aprendizaje profundo aplicadas al pronóstico de la volatilidad realizada horaria del Bitcoin a partir de datos de alta frecuencia del mercado spot. Se implementaron y compararon cinco metodologías: el modelo Heterogeneous Autoregressive (HAR) como referencia econométrica, CatBoost como representante de gradient boosting sobre datos tabulares, y tres arquitecturas de redes neuronales profundas (LSTM, GRU y Transformer). La variable objetivo se construyó como el logaritmo de la volatilidad realizada horaria adelantada un período. El conjunto de predictores incluyó componentes autorregresivos, proxies de volatilidad intraperiodo, variables de microestructura de mercado y codificación temporal cíclica. Los resultados fuera de muestra evaluados con RMSE, MAE y QLIKE sobre datos de julio de 2022 a abril de 2026 indican que CatBoost alcanzó el mejor desempeño global, seguido por GRU y HAR. Estos hallazgos sugieren que, bajo las condiciones analizadas, una ingeniería de características tabular bien construida combinada con métodos de gradiente no lineal supera a las arquitecturas secuenciales profundas, incluso cuando estas últimas disponen de ventanas temporales amplias. El estudio aporta a la literatura sobre pronóstico de volatilidad en criptoactivos evidencia comparativa sistemática bajo condiciones de alta frecuencia, y subraya la necesidad de confrontar arquitecturas complejas contra líneas base parsimoniosas antes de adoptarlas en aplicaciones de riesgo financiero.

Palabras clave: Volatilidad realizada, Bitcoin, aprendizaje profundo, pronóstico de series temporales financieras

Abstract

Bitcoin volatility presents one of the most persistent challenges in risk management within crypto-asset markets: its discontinuous nature, extreme sensitivity to news, and nonlinear depen-

dence structure mean that no single model definitively captures its behavior. This paper examines the predictive capacity of distinct econometric, machine learning, and deep learning approaches applied to forecasting the hourly realized volatility of Bitcoin using high-frequency spot market data. Five methodologies were implemented and compared: the Heterogeneous Autoregressive (HAR) model as the econometric benchmark, CatBoost as a representative of gradient boosting on tabular data, and three deep neural network architectures (LSTM, GRU, and Transformer). The target variable was constructed as the logarithm of one-period-ahead hourly realized volatility. The predictor set included autoregressive components, intra-period volatility proxies, market microstructure variables, and cyclical time encoding. Out-of-sample results evaluated with RMSE, MAE, and QLIKE over data spanning July 2022 to April 2026 indicate that CatBoost achieved the best overall performance, followed by GRU and HAR. These findings suggest that, under the conditions analyzed, well-engineered tabular features combined with nonlinear gradient methods outperform deep sequential architectures, even when the latter have access to extended temporal windows. The study contributes to the literature on volatility forecasting in crypto-asset markets by providing systematic comparative evidence under high-frequency conditions and underscores the need to benchmark complex architectures against parsimonious baselines before adopting them in financial risk applications.

Keywords: Realized volatility, Bitcoin, deep learning, financial time series forecasting

1 Introducción

En las dos últimas décadas, BTC (Bitcoin, por sus siglas en inglés) se ha consolidado como el activo digital más representativo del ecosistema de criptomonedas, y ha funcionado también como un laboratorio natural para el análisis de series temporales financieras de alta complejidad. A diferencia de otros mercados, el BTC opera las veinticuatro horas, los siete días de la semana, sin cierres ni pausas institucionales. Esta característica, combinada con niveles de volatilidad muy superiores a los de activos tradicionales, una marcada sensibilidad a choques exógenos y una dinámica dominada por relaciones no lineales, dificulta la identificación de patrones estables y hace del pronóstico de volatilidad un problema de especial interés tanto en la investigación como en la práctica profesional. Una mayor precisión predictiva se traduce directamente en mejores estimaciones de VaR (Value at Risk, por sus siglas en inglés), Expected Shortfall y otras medidas que sustentan decisiones de inversión, cobertura y diseño de estrategias cuantitativas (Baek & Elbeck, 2015; Gkillas & Katsiampa, 2018; Katsiampa, 2017).

Desde una perspectiva metodológica, el Bitcoin impone condiciones que no son fáciles de satisfacer. Opera en un entorno digital descentralizado, intensamente especulativo y expuesto a cambios

abruptos en liquidez, sentimiento de mercado y actividad on-chain. No exhibe los patrones de cierre-apertura que caracterizan a los mercados convencionales, y responde con rapidez a eventos informativos y regulatorios (Katsiampa, 2017; Baur et al., 2018). A esto se suma una microestructura con fragmentación marcada de la liquidez entre múltiples exchanges, variabilidad en los diferenciales de compra y venta, y la potencial presencia de operaciones de wash trading (Griffin & Shams, 2020; Makarov & Schoar, 2020). En conjunto, estos factores limitan la utilidad de enfoques estadísticos tradicionales como línea única de análisis. Modelos como ARIMA (AutoRegressive Integrated Moving Average, por sus siglas en inglés), GARCH (Generalized Autoregressive Conditional Heteroskedasticity, por sus siglas en inglés, o Heterocedasticidad Condicional Autorregresiva Generalizada) y sus extensiones han sido ampliamente empleados para predecir volatilidad, pero su capacidad se ve restringida cuando la señal contiene dependencias no lineales, ruido elevado y estructuras dinámicas que cambian rápidamente, rasgos que definen al mercado de criptomonedas (Engle, 1982; Bollerslev, 1986; Hansen & Lunde, 2005).

En el contexto de datos de alta frecuencia, la volatilidad realizada ha emergido como una proxy ex-post robusta de la varianza verdadera no observable. Definida como la suma de los

retornos al cuadrado en un período dado, constituye un estimador no paramétrico de la varianza integrada que aprovecha la disponibilidad de datos tick-by-tick (Andersen & Bollerslev, 1998; Barndorff-Nielsen & Shephard, 2002). Para el Bitcoin, la disponibilidad de datos a muy alta frecuencia permite construir medidas de volatilidad realizada a intervalos que van desde el minuto hasta la hora (Katsiampa et al., 2019). El modelo HAR (Heterogeneous Autoregressive, por sus siglas en inglés) de volatilidad realizada, propuesto por Corsi (2009), ha ganado amplia aceptación como punto de referencia gracias a su capacidad para integrar información de distintos horizontes temporales en una estructura parsimoniosa, con documentada robustez en criptomonedas (Corsi et al., 2008; Wang et al., 2020). Sin embargo, su dependencia de relaciones funcionales lineales o paramétricas rígidas puede impedirle capturar las no linealidades y cambios de régimen propios de mercados altamente volátiles (Bauwens et al., 2006).

Frente a estas limitaciones, los modelos de aprendizaje profundo han tomado protagonismo porque aprenden representaciones complejas directamente desde los datos y modelan dependencias temporales de distinto alcance con menos supuestos sobre la forma funcional de la serie. Los métodos de gradient boosting XGBoost (Extreme Gradient Boosting, por sus siglas en inglés), LightGBM (Light Gradient Boosting Machine, por sus siglas en inglés) y CatBoost destacan por identificar relaciones no lineales complejas en datos tabulares mediante la construcción iterativa de ensambles de árboles débiles (Ke et al., 2017; Prokhorenkova et al., 2018). Las redes LSTM (Long Short-Term Memory, por sus siglas en inglés) y GRU (Gated Recurrent Unit, por sus siglas en inglés), por su parte, están diseñadas para capturar dependencias de largo plazo en secuencias

temporales, atenuando el problema del gradiente evanescente que afecta a las redes recurrentes estándar (Hochreiter & Schmidhuber, 1997; Cho et al., 2014). En el mercado cripto, arquitecturas recurrentes, convolucionales e híbridas han mostrado resultados prometedores: Shen et al. (2021) documentaron que redes LSTM superan a modelos GARCH en el pronóstico de volatilidad diaria de Bitcoin, mientras que (Livieris et al., 2021) propusieron una arquitectura híbrida CNN-LSTM (Convolutional Neural Network–Long Short-Term Memory, por sus siglas en inglés) que reduce el error de predicción respecto a modelos individuales.

Evidencia más reciente extiende estos hallazgos hacia escenarios operativos. (Tang et al., 2023) calibraron una red GRU mediante optimización por enjambre de partículas para el pronóstico de volatilidad de futuros de Bitcoin, reportando superioridad frente a GARCH-MIDAS (Generalized Autoregressive Conditional Heteroskedasticity–Mixed Data Sampling, por sus siglas en inglés). (Bucci, 2020) demostró que incorporar variables de microestructura mejora

sustancialmente el ajuste de modelos de volatilidad realizada. Estudios comparativos más amplios, como Kim et al. (2025), encontraron que XGBoost supera a LSTM en pronóstico de volatilidad de futuros de petróleo cuando se dispone de características bien diseñadas, lo que sugiere que la elección del modelo depende del contexto y de la calidad del espacio de características, no de la complejidad arquitectónica en sí. En cuanto a los Transformers basados en mecanismos de atención multi-cabeza y procesamiento paralelo de secuencias (Vaswani et al., 2017) Herremans y Low (2022) los exploraron para clasificación de eventos extremos de volatilidad en criptomonedas, aunque el debate sobre su ventaja comparativa frente a redes recurrentes en pronóstico continuo de volatilidad de Bitcoin sigue abierto y con resultados mixtos (Li et al., 2019; Zhou et al., 2021; Lim & Zohren, 2021).

La literatura, sin embargo, todavía presenta vacíos relevantes. Los estudios difieren en la variable objetivo, la frecuencia temporal, la longitud de ventana, los predictores utilizados y los criterios de partición entre entrenamiento y prueba, lo que dificulta comparaciones directas y conclusiones generalizables. La mayoría se concentra en frecuencias diarias o semanales, dejando relativamente inexploradas las frecuencias horarias, que son críticas para el trading algorítmico y la gestión de riesgo intradiario. Las comparaciones sistemáticas entre gradient boosting, redes recurrentes y Transformers bajo un marco experimental homogéneo son escasas, y la evidencia sobre la efectividad de los Transformers para pronóstico continuo de volatilidad de Bitcoin permanece fragmentada. Quizás la pregunta más relevante para la práctica sigue sin respuesta definitiva: ¿la complejidad adicional de las arquitecturas profundas se traduce en ganancias predictivas sustanciales cuando el espacio de características ha sido bien diseñado? (McAleer & Medeiros, 2011; Lim & Zohren, 2021).

Todo lo anterior se vuelve aún más relevante cuando dejamos de trabajar con datos de cierre

diario. La frecuencia a la que se miden los retornos importa: no es lo mismo analizar el Bitcoin hora por hora que día a día. A mayor frecuencia, más ruido hay en los datos, los patrones son menos estables y el modelo tiene que procesar mucha más información. Por eso, la literatura recomienda agregar los retornos en intervalos de entre 5 y 15 minutos —ni tan cortos que el ruido domine, ni tan largos que se pierda información valiosa (Andersen et al., 2001; Bandi & Russell, 2006). Sin embargo, pocos estudios han comparado de forma rigurosa distintos modelos de aprendizaje profundo y gradient boosting trabajando específicamente con volatilidad horaria del Bitcoin, bajo las mismas condiciones experimentales. Ese es precisamente el vacío que este trabajo busca cubrir.

El presente trabajo se sitúa en la intersección entre aprendizaje profundo, series temporales financieras y gestión cuantitativa del riesgo. El objetivo central es comparar el desempeño predictivo de cinco metodologías HAR, CatBoost, LSTM, GRU y Transformer en el pronóstico de volatilidad realizada horaria del Bitcoin, bajo un esquema experimental unificado de preprocesamiento, ingeniería de características, entrenamiento y validación temporal. El interés no radica únicamente en minimizar el error de predicción, sino en evaluar qué tipo de metodología captura con mayor eficacia la dinámica del activo cuando se cuenta con un espacio de características bien diseñado, y en qué medida la sofisticación arquitectónica genera ventajas predictivas reales frente a enfoques más parsimoniosos. La hipótesis central es que, cuando se dispone de características tabulares bien diseñadas que incorporan información autorregresiva, proxies de volatilidad intraperiodo, variables de microestructura y codificación temporal cíclica los métodos de gradient boosting pueden alcanzar un desempeño competitivo o superior al de las arquitecturas profundas secuenciales, dado que explotan eficientemente las interacciones no lineales sin el costo computacional asociado al modelado de representaciones secuenciales complejas.

La relevancia de esta investigación opera en dos planos. En el académico, contribuye al debate sobre la idoneidad de arquitecturas profundas frente a métodos tabulares en problemas financieros no lineales, bajo un marco experimental homogéneo que permite comparaciones directas. Documentar empíricamente las ventajas relativas del gradient boosting sobre características bien diseñadas, validar la robustez de líneas base econométricas como HAR, y señalar las limitaciones de arquitecturas Transformer en contextos de datos limitados y dependencias locales constituyen aportes que la literatura aún necesita consolidar. En el plano aplicado, los resultados tienen implicaciones concretas para profesionales en finanzas cuantitativas que deben tomar decisiones sobre qué modelos adoptar para gestión de riesgo intradiario.

2 Metodología

2.1 Descripción de los datos

El estudio se construyó a partir de datos históricos descargados mediante la API (Application Programming Interface, por sus siglas en inglés) de Binance. La fuente primaria correspondió a velas de cinco minutos disponibles desde 2017, con información de apertura, máximo, mínimo, cierre, volumen transado y volumen comprador agresivo; este último se utilizó para construir el Taker Buy Ratio. Aunque la granularidad original de los datos es de 5 minutos, el problema de modelado se formuló a frecuencia horaria. Sobre esta estructura, la variable objetivo y_t quedó definida como la volatilidad realizada logarítmica adelantada un período:

$$y_t = \ln(RV_{t+1}) \quad (1)$$

Esta formulación es consistente con la literatura que recomienda trabajar con volatilidad realizada y transformaciones logarítmicas para capturar persistencia, heterogeneidad temporal y relaciones no lineales en problemas de pronóstico financiero (Bucci, 2020; McAleer & Medeiros, 2011; Shen et al., 2021).

En este orden de ideas, para llegar a la volatilidad realizada logarítmica, primero se calcula una variable proxy de la volatilidad realizada RV_h , la cual se construyó a partir de los precios de cierre de 5 minutos. En primer lugar, se aplicó logaritmo natural a la serie de cierres x_t , de modo que

$$x_t = \ln(C_t). \quad (2)$$

donde C_t representa el precio de cierre observado en el instante t . Seguidamente, los retornos logarítmicos intraperiodo r_t se calcularon por diferenciación de primer orden:

$$r_t = x_t - x_{t-1}. \quad (3)$$

Finalmente, cada retorno logarítmico se elevó al cuadrado y se remuestreó a frecuencia horaria mediante suma, de manera que:

$$RV_h = \sum_{i=1}^{12} r_{h,i}^2. \quad (4)$$

donde cada hora reúne doce intervalos de cinco minutos. Finalmente, para estabilizar la distribución de la serie y facilitar el modelado, se aplicó una transformación logarítmica adicional sobre la

medida agregada, obteniendo $\ln(RV_h)$. Esta serie transformada fue la base sobre la que se construyó el objetivo adelantado empleado en la etapa predictiva.

La Figura 1 resume visualmente la secuencia de transformación de la serie original, mostrando los precios de cierre crudos, los retornos logarítmicos y la volatilidad realizada en escala logarítmica empleada en el modelado.

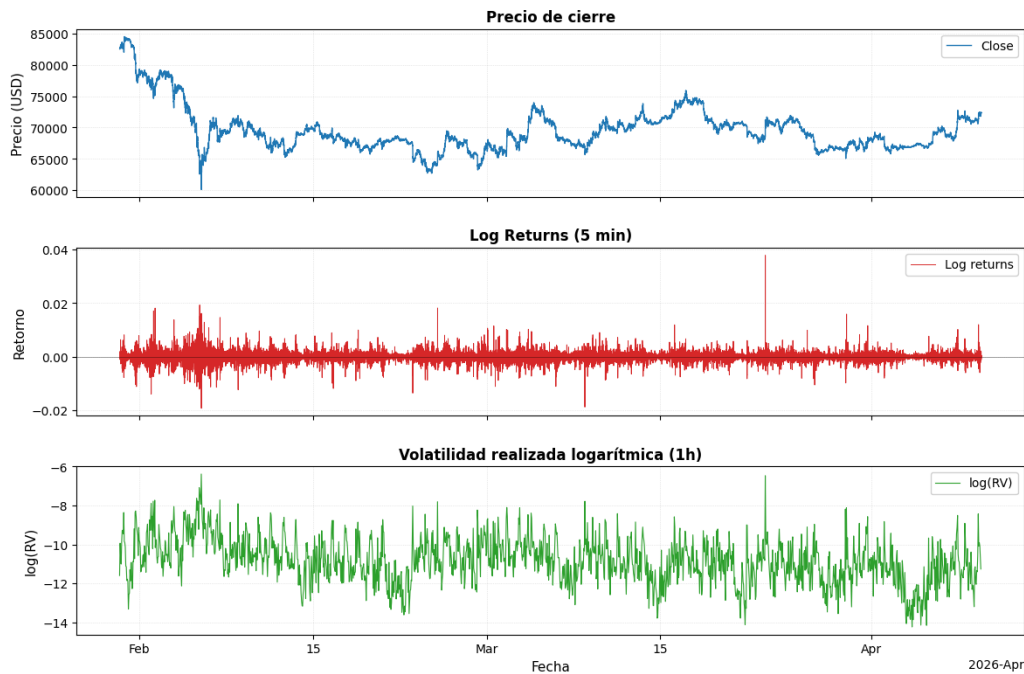


Figura 1

Precio de cierre, log returns y volatilidad realizada logarítmica.

La matriz de predictores se construyó a partir de cinco familias de variables: (i) componentes autorregresivos y de memoria sobre $\ln(RV)$ (Realized Volatility, por sus siglas en inglés) y sobre los retornos logarítmicos en rezagos de 1, 2, 3, 6 y 12 períodos; (ii) proxies de volatilidad intraperiodo basados en los spreads $\ln(High/Low)$ y $\ln(Close/Open)$; (iii) variables de microestructura y liquidez, incluyendo $\ln(Volume + 1)$, sus medias móviles en ventanas de 12 y 48 períodos, y el Taker Buy Ratio; (iv) características móviles de volatilidad histórica y momento acumulado; y (v) codificación cíclica de la hora del día mediante funciones seno y coseno. A nivel exploratorio, la serie mostró persistencia temporal, agrupamiento de volatilidad, episodios de variación abrupta y estacionalidad intradiaria, patrones que justificaron el uso de rezagos, estadísticas móviles, proxies de rango y codificación temporal cíclica.

2.2 Preprocesamiento y limpieza de los datos

El preprocesamiento tuvo como propósito convertir una serie de mercado en una matriz tabular y secuencial apta para los distintos tipos de modelos evaluados. La limpieza inicial contempló el tratamiento de valores nulos mediante propagación hacia adelante (forward fill), decisión que permitió conservar la continuidad temporal de la serie antes de construir las variables derivadas. Sobre la información depurada se aplicó la secuencia de transformaciones descrita para obtener el proxy horario de volatilidad realizada y su versión logarítmica.

Una vez construida la serie objetivo, se generaron las variables explicativas mediante logaritización de covariables de escala, construcción de rezagos sobre precios, retornos y volatilidad, cálculo de medias y desviaciones estándar móviles, generación de indicadores de momento y codificación trigonométrica de la hora del día. La microestructura del mercado se resumió, entre otras variables, mediante la proporción de volumen comprador agresivo respecto al volumen total:

$$\text{Taker Buy Ratio}_t = \frac{V_{buy,t}}{V_t}. \quad (5)$$

y la componente temporal cíclica se representó como:

$$h_t^{(\sin)} = \sin\left(2\pi \frac{\text{hora}_t}{24}\right), \quad h_t^{(\cos)} = \cos\left(2\pi \frac{\text{hora}_t}{24}\right). \quad (6)$$

La construcción de rezagos y ventanas móviles implicó la eliminación operativa de observaciones iniciales sin soporte completo para todas las variables derivadas. La partición de los datos se realizó de forma cronológica para evitar fuga de información entre periodos. El conjunto efectivo empleado para modelado quedó delimitado entre el 13 de julio de 2022 y el 9 de abril de 2026, con la siguiente distribución temporal:

Tabla 1

Partición cronológica de los datos.

Conjunto	Inicio	Fin
Entrenamiento	2022-07-13 00:00:00	2025-03-04 04:00:00
Validación	2025-03-04 05:00:00	2025-09-26 23:00:00
Prueba	2025-09-27 00:00:00	2026-04-09 19:00:00

2.3 Método de análisis y modelado

El diseño experimental comparó modelos de distinta complejidad para establecer una referencia interpretable y evaluar la ganancia potencial derivada de técnicas no lineales. En primer lugar, se implementó un modelo HAR de volatilidad realizada, construido con componentes de corto, mediano y mayor alcance mediante rezagos de 1, 12 y 78 periodos. Su especificación puede resumirse como:

$$\hat{y}_{t+1} = \beta_0 + \beta_1 \ln(RV_t) + \beta_2 \ln(RV)_{t,12} + \beta_3 \ln(RV)_{t,78} + \varepsilon_{t+1}. \quad (7)$$

Donde $\ln(RV)_{t,12}$ y $\ln(RV)_{t,78}$ representan promedios móviles de 12 y 78 periodos, respectivamente. Este modelo se adoptó como referencia por su uso extendido en pronóstico de volatilidad realizada y por su capacidad para resumir la persistencia temporal con una estructura parsimoniosa (Corsi, 2009; McAleer & Medeiros, 2011).

En segundo lugar, se empleó CatBoost como representante de los métodos de gradient boosting sobre datos tabulares. En tercer lugar, se entrenaron tres arquitecturas profundas orientadas al modelado secuencial: LSTM, GRU y un Transformer de tipo encoder-only. Para estas últimas se utilizó una ventana de contexto de 128 observaciones, suficiente para incorporar memoria temporal y capturar persistencia y cambios de régimen. La elección de

estas familias fue coherente con la literatura reciente sobre volatilidad de Bitcoin, que destaca la utilidad de combinar baselines econométricos con modelos no lineales y secuenciales cuando se cuenta con un espacio de características enriquecido (Shen et al., 2021; Tang et al., 2023; Tripathy et al., 2024).

2.4 Evaluación del modelo

La evaluación se realizó fuera de muestra sobre una partición temporal estricta. Para comparar los modelos se reportaron tres métricas complementarias: RMSE (Root Mean Squared Error, por sus siglas en inglés), MAE (Mean Absolute Error, por sus siglas en inglés) y QLIKE (Quasi-Likelihood, por sus siglas en inglés) como criterio especializado para pronóstico de volatilidad. La validación privilegió la separación cronológica por encima de esquemas aleatorios, preservando la causalidad temporal y evitando estimaciones artificialmente optimistas. La comparación entre modelos se efectuó sobre el mismo horizonte de evaluación, garantizando una base homogénea para interpretar las diferencias observadas.

3 Resultados y Discusión

La evaluación fuera de muestra sobre el conjunto de prueba comprendido entre septiembre de 2025 y abril de 2026 reveló diferencias claras entre los cinco modelos. La Tabla 1 presenta los resultados bajo tres métricas complementarias: RMSE y MAE como medidas de error en magnitud, y QLIKE como criterio especializado que penaliza asimétricamente la subestimación.

CatBoost obtuvo el mejor desempeño global, con mejoras de 7.1 % en RMSE y 9.6 % en MAE respecto al segundo mejor modelo, GRU. Este resultado merece atención: CatBoost opera sobre características tabulares precomputadas, mientras que GRU procesa secuencias temporales de 128 observaciones mediante arquitecturas recurrentes. Que un modelo tabular supere a uno secuencial en este contexto no es trivial, y sugiere algo importante sobre la naturaleza del problema. Cuando el espacio de características está bien construido incorporando rezagos, ventanas móviles, proxies de volatilidad intraperiodo, variables de microestructura y codificación temporal cíclica el gradient boosting puede explotar las

interacciones no lineales de manera más eficiente que el aprendizaje de representaciones latentes secuenciales.

Tabla 2

Desempeño fuera de muestra de los modelos evaluados.

Modelo	RMSE	MAE	QLIKE
CatBoost	0.7917	0.5989	-9.6699
GRU	0.8525	0.6539	-9.5207
HAR	0.8608	0.6625	-9.6214
LSTM	0.8733	0.6697	-9.4988
Transformer	0.9375	0.7415	-9.6831

Este hallazgo es coherente con evidencia de las competencias de pronóstico M4 y M5, donde el gradient boosting sobre características diseñadas ha mostrado desempeño competitivo o superior en escenarios tabulares estructurados y ventanas temporales no extremadamente largas (Makridakis et al., 2018, 2020; Kim et al., 2025). No obstante, esta ventaja es específica al contexto donde: (i) la ingeniería de características fue rigurosa, (ii) los datos de entrenamiento fueron de tamaño moderado, (iii) las dependencias no presentaron estructura de muy largo alcance, y (iv) la información predictiva ya estaba explícitamente codificada en los predictores. En contextos con características limitadas, dependencias irregulares de largo plazo o conjuntos masivos de datos, las arquitecturas profundas podrían ofrecer ventajas mediante el aprendizaje automático de representaciones.

GRU se posiciona en segundo lugar, muy cerca de HAR, y supera a LSTM. Este patrón confirma dos hallazgos relevantes. Primero, las arquitecturas recurrentes más simples con menos parámetros pueden favorecer la generalización en tamaños de muestra moderados, en línea con el principio de parsimonia (Chung et al., 2014; Tang et al., 2023). Segundo, y quizás más llamativo, la robustez de HAR como línea base econométrica es notable: con apenas cuatro parámetros estimados mediante regresión lineal sobre promedios móviles en tres escalas temporales, mantiene una capacidad predictiva competitiva frente a arquitecturas

mucho más complejas cuando la señal dominante proviene de la persistencia autorregresiva (Corsi, 2009). Desde una perspectiva aplicada, este resultado refuerza la conveniencia de establecer líneas base econométricas sólidas antes de invertir en arquitecturas que demandan mayor expertise, capacidad computacional y tiempo de desarrollo. Adicionalmente, HAR ofrece interpretabilidad directa, coeficientes legibles como elasticidades e inferencia clásica, ventajas que los modelos de gradient boosting y aprendizaje profundo solo pueden aproximar mediante técnicas especializadas de explicabilidad (Lundberg & Lee, 2017; Molnar, 2020).

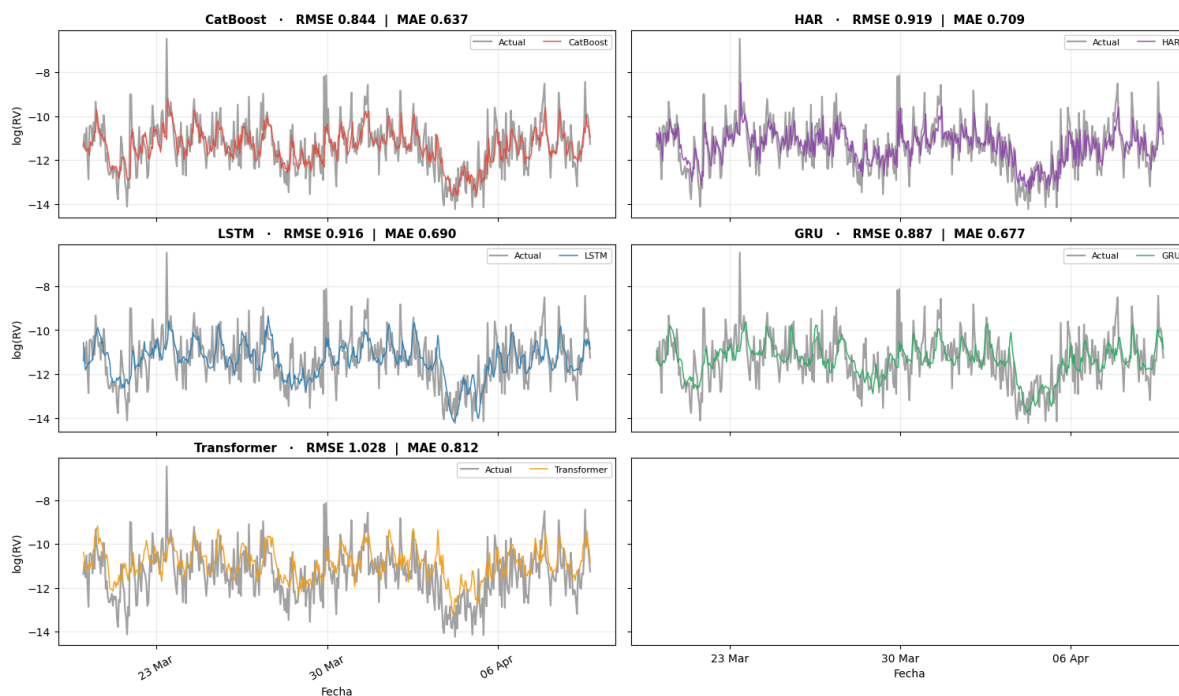


Figura 2

Predicciones en las últimas 500 horas del test set.

El Transformer exhibe el peor desempeño, con deterioros de 18.4 % en RMSE y 23.8 % en MAE respecto a CatBoost, resultado que contrasta con los hallazgos prometedores reportados en otras aplicaciones (Li et al., 2019; Zhou et al., 2021). Tres factores pueden explicar esta debilidad. Primero, los mecanismos de atención global pueden introducir más ruido que señal cuando la

dependencia temporal es predominantemente local y de corto-mediano plazo, como ocurre en volatilidad horaria donde la mayor parte de la información predictiva está

concentrada en rezagos recientes. Segundo, el entrenamiento efectivo de Transformers típicamente requiere millones de observaciones; la muestra disponible en este estudio es moderada, insuficiente para estimar robustamente todos los parámetros de atención multi-cabeza sin sobreajuste a patrones espurios. Tercero, la arquitectura vanilla empleada no incorpora adaptaciones específicas para series temporales atención sparse, descomposición tendencia-estacionalidad o embeddings posicionales mejorados presentes en variantes especializadas como Informer, Autoformer o Temporal Fusion Transformer (Zhou et al., 2021; Wu et al., 2021; Lim & Zohren, 2021). Este resultado advierte sobre la adopción acrítica de arquitecturas de moda: la sofisticación no reemplaza la necesidad de una relación señal-ruido favorable, características informativas y adecuación arquitectónica al problema específico (Makridakis et al., 2018).

Vale la pena detenerse en la sensibilidad del ordenamiento a la métrica utilizada. CatBoost domina en RMSE y MAE, pero su ventaja se reduce en QLIKE, donde HAR resulta prácticamente equivalente (QLIKE = -9.6699 frente a -9.6214). Además, LSTM muestra el peor QLIKE (-9.4988) a pesar de no ser el peor modelo en las otras métricas. Este patrón confirma que la comparación de modelos de volatilidad es sensible a la función de pérdida, y que el modelo con menor error promedio no siempre coincide con el de mejor comportamiento bajo criterios de riesgo que penalizan asimétricamente subestimación y sobrestimación (Patton, 2011; Hansen & Lunde, 2005). La implicación práctica es directa: la selección final debe considerar múltiples métricas alineadas con las preferencias de la aplicación específica QLIKE para gestión de riesgo con penalización severa de subestimación, MAE para señales de trading donde el error absoluto determina la rentabilidad esperada.

El análisis visual de las predicciones sobre las últimas 500 horas del conjunto de prueba complementa el análisis cuantitativo. CatBoost y GRU reproducen con mayor fidelidad las oscilaciones locales, capturando tanto picos de volatilidad como retornos a niveles bajos, sin suavización excesiva ni retardo temporal pronunciado. HAR conserva un comportamiento estable, capturando tendencias de mediano plazo, pero con mayor suavización, esperable dada su estructura basada en promedios móviles. LSTM muestra capacidad intermedia para

seguir cambios de nivel, con menor suavización que HAR pero menor adaptabilidad local que CatBoost y GRU. El Transformer evidencia la separación más pronunciada respecto a los valores observados, especialmente en segmentos de volatilidad cambiante, con un sesgo aparente hacia la media incondicional y dificultad para ajustarse rápidamente ante regímenes no observados durante el entrenamiento.

En su conjunto, la evidencia complementa y extiende la literatura existente. Shen et al. (2021) reportan que LSTM supera a GARCH en volatilidad diaria; el cuarto lugar de LSTM en este estudio sugiere que esa ventaja no es uniforme y depende de la frecuencia, del diseño de características y del periodo de evaluación. Tang et al. (2023) encuentran que una GRU optimizada supera a GARCH-MIDAS, coherente con el segundo lugar de GRU aquí. Kim et al. (2025) reportan que XGBoost supera a LSTM en futuros de petróleo, consistente con la robustez de CatBoost frente al aprendizaje profundo. Bucci (2020) demuestra que las variables de microestructura mejoran el ajuste de modelos de volatilidad realizada; el diseño experimental de este trabajo incorporó volumen, Taker Buy Ratio y spreads, confirmando su valor predictivo a través del desempeño de CatBoost.

4 Conclusiones

Este estudio comparó cinco metodologías de pronóstico de volatilidad horaria de Bitcoin bajo un marco experimental homogéneo: HAR como línea base econométrica, CatBoost como representante de gradient boosting tabular, y tres arquitecturas de aprendizaje profundo LSTM, GRU y Transformer. A partir de datos de velas de 5 minutos agregadas a nivel horario desde 2017, con predictores que incorporan componentes autorregresivos, proxies de volatilidad intraperiodo, variables de microestructura y codificación temporal cíclica, la evaluación fuera de muestra revela que CatBoost obtiene el mejor desempeño global ($RMSE = 0.7917$, $MAE = 0.5989$), con mejoras de entre 7 % y 10 % respecto al segundo mejor modelo (GRU). Este resultado pone de manifiesto que, cuando el espacio de características está bien diseñado mediante ingeniería informada por el dominio, la combinación de predictores tabulares con gradient boosting no lineal puede superar a arquitecturas profundas secuenciales incluso cuando estas disponen de ventanas temporales amplias. La sofisticación arquitectónica, en otras palabras, no sustituye la calidad de las características predictivas.

GRU se posiciona en segundo lugar, muy cerca de HAR, confirmando que una línea base econométrica bien especificada es difícil de desplazar cuando la persistencia autorregresiva constituye la señal dominante. El superior desempeño de GRU frente a LSTM refuerza el principio de parsimonia: en tamaños de muestra moderados, la simplificación arquitectónica tiende a mejorar la generalización. El Transformer, en el extremo opuesto, exhibe el peor desempeño del grupo ($RMSE = 0.9375$, $MAE = 0.7415$), con deterioros de entre 18 % y 24 % respecto a CatBoost, un resultado que contrasta con los reportados en otras aplicaciones y que advierte que su adopción para pronóstico de volatilidad debe respaldarse en evidencia empírica sólida sobre el contexto específico. La sensibilidad del ordenamiento a la métrica de evaluación es igualmente relevante: CatBoost domina en RMSE y MAE, pero su ventaja se reduce en QLIKE, donde HAR resulta prácticamente equivalente, lo que subraya la necesidad de evaluar múltiples criterios según las preferencias de

pérdida de cada aplicación práctica.

En el plano de los aportes, la investigación provee evidencia empírica rigurosa sobre el desempeño comparativo de modelos bajo datos de alta frecuencia horaria, complementando estudios concentrados en frecuencias diarias y entregando puntos de referencia cuantitativos para trading algorítmico y gestión de riesgo intradiario. Confirma que el gradient boosting sobre características bien diseñadas es una alternativa robusta frente al aprendizaje profundo, valida la robustez de líneas base econométricas como HAR, y alerta sobre las limitaciones de los Transformers vanilla en contextos de datos limitados y dependencias predominantemente locales. Para profesionales en finanzas cuantitativas, los resultados tienen implicaciones concretas: establecer líneas base sólidas antes de invertir en arquitecturas complejas, priorizar la ingeniería de características que capture microestructura y estacionalidad, evaluar múltiples métricas alineadas con la función de pérdida implícita, preservar la causalidad temporal en la validación, privilegiar la interpretabilidad en contextos regulatorios, y considerar extensiones probabilísticas para la caracterización de riesgo de cola VaR y expected shortfall.

El estudio reconoce limitaciones que deben tenerse presentes. La evidencia se basa en un único criptoactivo (Bitcoin), un único exchange (Binance) y una única frecuencia (horaria), lo que acota la generalización de los hallazgos. El enfoque se limita a predicción puntual, sin abordar intervalos probabilísticos ni cuantificación de incertidumbre. Los predictores no incorporan variables exógenas enriquecidas métricas on-chain, sentimiento, eventos regulatorios o factores macroeconómicos cuya inclusión podría modificar el ordenamiento entre modelos, especialmente beneficiando a las arquitecturas profundas. No se implementaron técnicas de interpretabilidad como SHAP (SHapley Additive exPlanations, por sus siglas en inglés) o permutation importance, ni se exploró validación mediante ventanas rodantes. El Transformer evaluado corresponde a la arquitectura vanilla sin adaptaciones para series temporales, lo que puede explicar parcialmente su bajo desempeño.

Futuras investigaciones podrán incorporar variables exógenas enriquecidas, explorar horizontes múltiples de pronóstico, desarrollar enfoques probabilísticos con densidad condicional, implementar técnicas de interpretabilidad, evaluar variantes especializadas de Transformers con mayor volumen de datos y regularización agresiva, y validar los hallazgos sobre una muestra amplia de criptoactivos y exchanges. La combinación de ingeniería de características rigurosa, líneas base econométricas sólidas y evaluación experimental homogénea sigue siendo el camino más prometedor para avanzar tanto en la comprensión teórica como en la utilidad práctica del pronóstico de volatilidad en mercados digitales.

5 Referencias

- Ahmed, A. (2025). Explainable Deep Learning for Financial Volatility Forecasting: An LSTM-Attention-SHAP Framework with Comprehensive Validation. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.5101234>
- Andersen, T. G., & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, 39(4), 885–905.
- Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics*, 61(1), 43–76.
- Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, 71(2), 579–625.
- Andersen, T. G., Bollerslev, T., & Diebold, F. X. (2006). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, 89(4), 701–720.
- Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79.
- Baek, C., & Elbeck, M. (2015). Bitcoins as an investment or speculative vehicle? A first look. *Applied Economics Letters*, 22(1), 30–34.
- Bandi, F. M., & Russell, J. R. (2006). Separating microstructure noise from volatility. *Journal of Financial Economics*, 79(3), 655–692.
- Barndorff-Nielsen, O. E., & Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society: Series B*, 64(2), 253–280.
- Baur, D. G., Hong, K., & Lee, A. D. (2018). Bitcoin: Medium of exchange or speculative assets? *Journal of International Financial Markets, Institutions and Money*, 54, 177–189.
- Bauwens, L., Laurent, S., & Rombouts, J. V. (2006). Multivariate GARCH models: A survey. *Journal of Applied Econometrics*, 21(1), 79–109.
- Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213.

-
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327.
- Bollerslev, T., Patton, A. J., & Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192(1), 1–18.
- Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, 18(3), 502–531. <https://doi.org/10.1093/jjfinec/nbaa008>
- Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.
- Corbet, S., Lucey, B., & Yarovaya, L. (2019). Datestamping the Bitcoin and Ethereum bubbles. *Finance Research Letters*, 26, 81–88.
- Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2), 174–196.
- Corsi, F., Mitnik, S., Pigorsch, C., & Pigorsch, U. (2008). The volatility of realized volatility. *Econometric Reviews*, 27(1–3), 46–78. <https://doi.org/10.1080/07474930701853616>
- Easley, D., López de Prado, M. M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *The Review of Financial Studies*, 25(5), 1457–1493.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007.
- Engle, R. F. (2001). GARCH 101: The use of ARCH/GARCH models in applied econometrics. *Journal of Economic Perspectives*, 15(4), 157–168.
- Gkillas, K., & Katsiampa, P. (2018). An application of extreme value theory to cryptocurrencies. *Economics Letters*, 164, 109–111.
- Griffin, J. M., & Shams, A. (2020). Is Bitcoin really untethered? *The Journal of Finance*, 75(4),

1913– 1964.

Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH (1,1)? *Journal of Applied Econometrics*, 20(7), 873–889.

Herremans, D., & Low, K. W. (2022). Forecasting Bitcoin volatility spikes from whale transactions and CryptoQuant data using Synthesizer Transformer models. *arXiv preprint arXiv:2209.03964*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy.

International Journal of Forecasting, 22(4), 679–688.

Katsiampa, P. (2017). Volatility estimation for Bitcoin: A comparison of GARCH models. *Economics Letters*, 158, 3–6.

Katsiampa, P., Corbet, S., & Lucey, B. (2019). High frequency volatility co-movements in cryptocurrency markets. *Journal of International Financial Markets, Institutions and Money*, 62, 35–52.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.

Kim, H. Y., Won, C. H., & Park, J. (2025). XGBoost vs. LSTM for oil futures volatility forecasting.

Energy Economics, 123, 106745. <https://doi.org/10.1080/00036846.2025.2458248>

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y. X., & Yan, X. (2019). Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in Neural Information Processing Systems*, 32, 5243–5253.

Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: A survey. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200209.

Livieris, I. E., Kiriakidou, N., Stavroyiannis, S., & Pintelas, P. (2021). An advanced CNN-LSTM model for cryptocurrency forecasting. *Electronics*, 10(3), 287.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances*

- in *Neural Information Processing Systems*, 30, 4765–4774.
- Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 135(2), 293–319.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4), 802–808.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74.
- Mandelbrot, B. (1963). The variation of certain speculative prices. *The Journal of Business*, 36(4), 394–419.
- McAleer, M., & Medeiros, M. C. (2011). Forecasting realized volatility with linear and nonlinear univariate models. *Journal of Economic Surveys*, 25(1), 6–18. <https://doi.org/10.1111/j.1467-6419.2010.00640.x>
- Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com.
- Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, 53(1), 61–65.
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256.
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31, 6638–6648.
- Shen, Z., Wan, Q., & Leatham, D. J. (2021). Bitcoin return volatility forecasting: A comparative study between GARCH and RNN. *Journal of Risk and Financial Management*, 14(7), 337. <https://doi.org/10.3390/jrfm14070337>
- Tang, Y., Song, Z., & Zhu, Y. (2023). An optimized GRU model for Bitcoin volatility forecasting with particle swarm optimization. *Computational Economics*. <https://doi.org/10.1007/s10614-023-10392-5>
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review.

International Journal of Forecasting, 16(4), 437–450.

Tripathy, N., et al. (2024). Bitcoin volatility forecasting: A comparative analysis of conventional econometric models with deep learning models. *International Journal of Power Electronics and Drive Systems*, 15(1), 614–623. <https://doi.org/10.11591/ijece.v15i1.pp614-623>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

Wang, J., Ma, F., Bouri, E., & Guo, Y. (2020). Which factors drive Bitcoin volatility: Macroeconomic, technical, or both? *Journal of Forecasting*, 39(8), 1281–1297.

Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with autocorrelation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34, 22419–22430.

Zahid, M., Iqbal, F., & Koutmos, D. (2022). Forecasting Bitcoin volatility using hybrid GARCH models with machine learning. *Risks*, 10(12), 237. doi:10.3390/risks10120237

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12), 11106–11115.