Market Efficiency of Bitcoin in 2024: Evidence from Statistical and Deep Learning Models

Donghua (Alex) Zheng

April 2025

Student ID: u2128931
Word Count: 4,881

Code Repository: github.com/alexzheng123/EC331-project

Acknowledgements

I would like to thank Dr Jeisson Cárdenas-Rubio, Dr Zeynep Kurter, and Dr Nathan Canen for their support and guidance in my journey leading to this study.

Abstract

This study examines the weak form efficient market hypothesis (EMH) in the Bitcoin market by analysing 1-minute trading data from 2024 using traditional statistical methods and deep learning models. Statistical tests such as the Ljung-Box test, Runs test, Variance Ratio test, Hurst exponent, and BDS test suggest significant autocorrelation in the return series, implying potential violations of weak form efficiency. To further investigate return predictability, the study employs ARIMA and GRU-based models, including a hybrid VMD-GRU-Attention model. The results show that the VMD-GRU-Attention model outperforms all other models by a wide margin in out-of-sample forecasting. These findings provide strong evidence against the weak form EMH in the 2024 Bitcoin market.

1 Introduction

Bitcoin (BTC), created by Nakamoto (2008), is a decentralised digital currency built on blockchain technology. It was designed to enable peer-to-peer transactions without reliance on banks or other intermediaries. Since its launch in 2009, Bitcoin has evolved from a niche experiment to a global financial asset. In 2021, El Salvador became the first country to adopt Bitcoin as legal tender. More recently, after the 2024 US presidential election, Bitcoin price has surged past $100,000 in December, pushing its market capitalisation above $2 trillion. This study uses the BTC/USD trading pair as a proxy for Bitcoin's market price.

Whether markets are efficient has been one of the most debated topics in finance and economics. Numerous studies have examined the efficiency of a wide range of assets in the financial industry. Fama (1970) introduced three forms of the efficient market hypothesis (EMH): the weak form, the semi-strong form, and the strong form. The weak form suggests that technical analysis that uses historical prices and trading volumes cannot be used to predict future returns or achieve consistent excess profits. The semi-strong form builds upon the weak form, stating that neither technical nor fundamental analysis can lead to consistent returns. The strong form argues that no public or private information can be used to generate excess profits. This study focuses on testing the weak form EMH.

Bitcoin has gained immense popularity among investors in recent years especially following the approval of Bitcoin exchange-traded funds (ETFs) in 2024. Since then, institutional investors have poured billions of dollars into the Bitcoin market. Unlike traditional financial assets such as stocks or forex, Bitcoin can be traded 24/7 on most cryptocurrency exchanges, with no central authority that controls its supply. These features make Bitcoin a unique case for studying market efficiency. Furthermore, understanding its efficiency not only helps traders refine their strategies but also allows regulators and policymakers to make more informed decisions in safeguarding the cryptocurrency market.

Whilst extensive research has examined Bitcoin's efficiency, most studies have relied on its daily prices. However, high-frequency intraday data contains much more information and algorithmic trading is now widespread in modern financial markets. Moreover, as more investors have entered the market, Bitcoin's efficiency may have evolved, necessitating updated analyses using more recent data. Despite the growing popularity of deep learning, a subfield of machine learning, and its ability to capture complex non-linear patterns in time series data that traditional models may fail to detect, few studies have directly compared deep learning models with traditional econometric approaches in assessing market efficiency. Therefore, deep learning models, such as the gated recurrent unit (GRU) network proposed by Chung et al. (2014) and the more sophisticated VMD-Attention-GRU (VMD-AttGRU) model, which was shown to be effective in forecasting stock price indices by Niu and Xu (2020), offer a novel approach to analysing the market efficiency of Bitcoin.

Inspired by the VMD-AttGRU model, this study proposes a VMD-GRU-Attention model that reverses the order of GRU and attention layers. This allows the attention layer to refine the outputs after the series has been processed by the GRU, thereby further improving forecast accuracy. This study aims to bridge these gaps by applying statistical tests, econometric, and deep learning models to assess Bitcoin's market efficiency using minute-level data.

This leads to the following research questions: Does the Bitcoin market exhibit weak form efficiency in 2024? And how do deep learning models perform compared to traditional econometric models in forecasting Bitcoin returns? In other words, the objectives of this study are to investigate the efficiency of the Bitcoin market and to compare the performance of deep learning and traditional econometric models in time series forecasting.

2 Literature Review

Urquhart (2016) conducted one of the earliest empirical studies on Bitcoin's weak form EMH (Fama, 1970) by applying a number of statistical tests including the autocorrelation, variance ratio, runs, BDS and Hurst exponent tests. Urquhart found that Bitcoin was inefficient over the full sample period but showed signs of becoming efficient over time. Extending this work, Nadarajah and Chu (2017) applied power transformations to Bitcoin returns and argued that the random walk hypothesis could not be rejected, thereby supporting weak form efficiency. Using overlapping windows to compute the Hurst exponent, Bariviera (2017) found that Bitcoin returns became more efficient over time, though persistent behaviour in daily returns was found between 2011 and 2014, reinforcing Urquhart's initial findings.

Building on the early literature, more recent studies found that Bitcoin's efficiency is not static but tends to evolve over time. Noda (2021) adopted a GLS-based time-varying autoregressive model and found that the degree of market efficiency varied with time, consistent with the adaptive market hypothesis (AMH) proposed by Lo (2004) which argues that market efficiency can emerge from time to time as investors adapt to changing conditions, including behavioural factors, structural changes, and external shocks. Kang et al. (2022) used random walk tests and an event study method to show that only 6.04% and 2.695% of the cryptocurrencies were weak form and semi-strong form efficient, respectively. Aslam et al. (2023) used the multifractal detrended fluctuation analysis (MFDFA) approach to study six major cryptocurrencies and found that Bitcoin and Litecoin exhibited the highest degrees of multifractality, i.e. market inefficiency. Although all cryptocurrencies displayed persistent or trending behaviour, their rolling window analysis showed an improvement in efficiency over time. In contrast, Yi et al. (2023) applied a quantum harmonic oscillator to test the weak form EMH in Bitcoin and concluded a near-efficient Bitcoin market. They found that Bitcoin was evolving into an efficient market and suggested that profitable opportunities for speculators may be increasingly limited.

The impact of external shocks on Bitcoin market efficiency has also been explored. Fernandes et al. (2022) examined five major cryptocurrencies by computing the permutation entropy and Fisher information measure (FIM), and found that their information efficiency remained stable before and after the COVID-19 outbreak. Kakinaka and Umeno (2022) investigated the effect of COVID-19 on efficiency across different time horizons and found that the pandemic reduced short-run efficiency whilst improving long-run efficiency.

To better capture Bitcoin's short-term behaviour, several studies have used intraday data to assess the weak form EMH. Aslan and Sensoy (2020) applied a battery of long memory tests to investigate the relationship between weak form efficiency and intraday sampling frequency. They found that the efficiency of cryptocurrencies followed a U-shaped pattern across sampling intervals: the 5-minute and 10-minute return series were weakly efficient whereas the 1-, 15-, and 60-minute returns were found to be informationally inefficient. De Nicola (2021) compared Bitcoin's intraday returns with those of traditional financial assets such as stocks and exchange rates, and found negative autocorrelation in 1-, 2-, and 4-hour returns, which was absent in the returns of other assets. This suggested a violation of the weak form EMH at the intraday level for the Bitcoin market.

Deep learning models, especially recurrent neural networks (RNNs), such as long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) and GRU (Chung et al., 2014) models, have gained popularity in the Bitcoin literature for their ability to capture complex non-linear patterns in time series data. Jaquart et al. (2021) applied a range of machine learning models on minutely Bitcoin prices to predict short-term movements in the Bitcoin market, with RNNs performing the best. The models yielded just above 50% accuracy. However, after factoring in transaction costs, trading results turned negative, suggesting the possibility of efficiency in the Bitcoin market. In a different study, Jaquart et al. (2022) compared various machine learning models trained on the 100 largest cryptocurrencies using live trading simulations, where they challenged the weak form EMH in the cryptocurrency market. Similarly, Goutte et al. (2023) used five years of hourly Bitcoin prices and technical indicators from 2017 to 2022 to compare several popular machine learning algorithms for trading signal generation. They found that RNNs performed the best, with the GRU model outperforming the LSTM model. Pourrezaee and Hajizadeh (2024) proposed stacking multiple machine learning models, including artificial neural network (ANN), support vector regression (SVR), LSTM, random forest, neural basis expansion analysis with exogenous variables (NBEATSx), the heterogeneous autoregressive (HAR) model, and GARCH, to forecast Bitcoin volatility and Value-at-Risk. Their results showed that the stacked model outperformed the individual models in terms of forecast error, and that using high-frequency data significantly improved predictive performance. Finally, Niu and Xu (2020) developed a hybrid model (VMD-AttGRU) that combined variational mode decomposition (VMD) and GRU with attention layers to forecast the FTSE 100 Index and the Nasdaq Composite Index. The VMD approach reduces the negative impact of noise in high-frequency data, and the attention layer allows the model to focus on the most relevant parts of the return series by assigning different weights. These changes make the already promising GRU network more effective at capturing hidden patterns in stock indices. They found that the hybrid model outperformed the individual models in predictive accuracy. These findings highlight the potential advantages of hybrid machine learning models and high-frequency data in improving the accuracy of time series forecasting.

Despite extensive literature on market efficiency and forecasting Bitcoin returns, there is limited research exploring the use of sophisticated hybrid deep learning models such as the VMD-AttGRU model to test the EMH using high-frequency intraday data. This is particularly surprising given that such models have been shown to outperform traditional econometric models in capturing complex non-linear patterns in time series data. Inspired by the VMD-AttGRU model, this study proposes a VMD-GRU-Attention model that reverses the order of GRU and attention layers. To bridge this gap, the paper applies a variety of statistical tests, the autoregressive integrated moving average (ARIMA) model, the GRU model, and the proposed VMD-GRU-Attention model, combined with 1-minute intraday data, to test whether the Bitcoin market follows the weak form EMH in 2024.

3 Data

In this paper, 1-minute BTC/USD open, high, low, close (OHLC) price and trading volume data from Bitstamp, covering the period from 1 January 2024 to 31 December 2024, are collected from CryptoDataDownload, totalling 3,056,184 observations. In addition, daily macroeconomic indicators, including the S&P 500 Index, Nasdaq 100 Index, Consumer Price Index (CPI), VIX Index, New York Gold Bullion price, and Brent Crude Oil price for the year 2024, are obtained from Finaeon, totalling 1,596 observations. Finally, 365 daily sentiment observations, as measured by the Crypto Fear & Greed Index, are collected from Alternative.me.

Macroeconomic indicators and sentiment data are used to extend the analysis beyond the weak form EMH and to explore the semi-strong form using a basic GRU network. The S&P 500 Index is included as a broad measure of the US economy, whilst the Nasdaq 100 Index is selected due to its strong correlation with Bitcoin close prices (Pearson correlation coefficient of 0.75 within the sample). The CPI captures inflation levels, and the VIX Index measures overall market volatility. The Crypto Fear & Greed Index gauges retail sentiment in the cryptocurrency market. These factors have been widely studied in the literature and are considered highly influential on Bitcoin prices (Huang et al., 2024; Bouteska et al., 2025; Chen, 2021; Huang, 2024).

Figure 1 shows Bitcoin prices and trading volumes in 2024. The price rose from just above \$40,000 to around \$100,000 by the end of December and appears to have reacted to major events annotated in the graph, such as the upward trend following the Bitcoin ETF approval in January and the sharp rally after Donald Trump's re-election in November. Trading volumes also spiked around these events. Whilst such reactions are common in traditional financial markets, they raise questions about the speed with which public information is incorporated into Bitcoin prices.

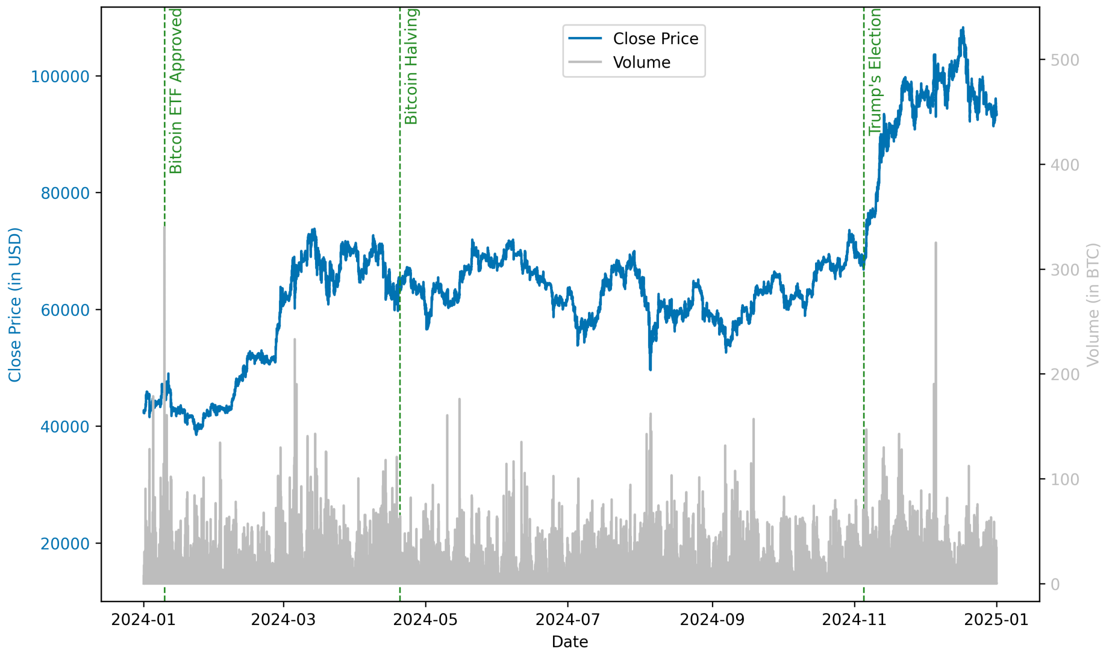


Figure 1: Bitcoin 1-minute close price and trading volume.

4 Methodology

4.1 Data Processing and Feature Generation

To prepare for empirical testing, minute-level logarithmic returns are computed using the close prices of Bitcoin. The log return at time t is calculated as shown in Equation 1, where P_t denotes the close price at time t:

$$R_t = \ln\left(\frac{P_t}{P_{t-1}}\right) \times 100 \tag{1}$$

Table 1 reports the descriptive statistics of 1-minute Bitcoin returns. The mean return is positive (0.0002) over the sample period, with a relatively large standard deviation (0.0756), indicating high volatility. The minimum and maximum returns deviate considerably from the mean, and the distribution exhibits excess kurtosis (56.7862) and negative skewness (-0.4262), consistent with Urquhart (2016). This suggests the presence of outliers and a higher likelihood of extreme negative returns relative to positive ones. The Jarque-Bera statistic further confirms the non-normality of Bitcoin returns, in line with Aslam et al. (2023).

Table 1: Descriptive statistics of 1-minute Bitcoin returns.

N	Mean	SD	Min	Max	Kurt	Skew	Jarque–Bera
509,363	0.0002	0.0756	-3.6120	2.6230	56.7862	-0.4262	6.8453E+07***

Notes: N denotes the number of observations. SD is the standard deviation. *** indicates significance at the 1% level.

In addition to raw OHLC prices and trading volumes, technical indicators are generated using the `TA-Lib` library in Python. These include the 20-period simple moving average (SMA), the 20-period rolling standard deviation as a proxy for volatility, and multiple momentum and trend indicators, including the relative strength index (RSI), the moving average convergence divergence (MACD), the rate of change (ROC), the commodity channel index (CCI), Williams% R, and the average direction index (ADX). These features are commonly used in technical analysis and quantitative trading.

4.2 Statistical Tests

In an efficient market, future prices are unpredictable and follow a random walk. To assess whether Bitcoin prices are consistent with the weak form EMH, statistical tests are applied to the log return series using the `statsmodels`, `arch`, and `hurst` libraries in Python.

First, the **Ljung-Box test** (Ljung and Box, 1978) is used to detect autocorrelation. The null hypothesis assumes no autocorrelation, so a significant result suggests that past returns may contain information about future returns, violating the weak form EMH. Second, the **Runs test** (Wald and Wolfowitz, 1940) is applied to test for randomness in Bitcoin's returns, under the null hypothesis that the series is random. Third, the **heteroscedasticity-robust Variance Ratio test** (Lo and MacKinlay, 1988) is employed to examine whether the return series follows a random walk. Under the null hypothesis, the variance of multi-period returns is proportional to the time interval. A significant result implies that the series is either mean-reverting or trending, both of which violate the weak form EMH.

Fourth, the **Hurst exponent** (H) is used to measure long-term memory in the series. According to Aslam et al. (2023), $H > 0.5$ indicates persistent or trending behaviour, whilst $H < 0.5$ suggests mean reversion. $H = 0.5$ implies that the series follows a purely random process. Fifth, a **rolling Hurst exponent** is computed using a moving window to analyse whether market efficiency evolves over time. Lastly, the **BDS test** is a non-parametric method that detects serial dependence and non-linear structure in time series data (Broock et al., 1996). A **rolling BDS test** is applied to examine whether the return series rejects the null hypothesis of an independently and identically distributed (i.i.d.) process.

4.3 Naïve Forecast Benchmark

To evaluate the performance of forecasting models, a naïve benchmark is constructed using the historical mean of past returns. Specifically, the mean return from the training set is used as a constant forecast for all time steps in both the validation and test sets. The predicted return $\hat{r}_t$ at any time t is defined in Equation 2, where $\bar{r}$ is the historical mean.

$$\hat{r}_t = \bar{r} \tag{2}$$

4.4 Autoregressive Integrated Moving Average Model

To examine whether historical returns can predict future returns, an autoregressive integrated moving average (ARIMA) model (Box and Jenkins, 1970) is applied to the log return series of Bitcoin. ARIMA models are widely used in financial forecasting for time series data.

An ARIMA(p, d, q) process consists of three components:

- **Autoregression (AR)**: regression of the variable on its own lagged values, of order p;
- **Integration (I)**: the number of differencing operations (d) required to make the series stationary;
- **Moving Average (MA)**: a linear combination of past forecast errors, of order q.

According to Kotu and Deshpande (2019), an ARIMA (p, d, q) process is defined in Equation 3, where the differenced series y'_t is modelled as a function of its lagged values and past forecast errors.

$$y'_t = I + \alpha_1 y'_{t-1} + \alpha_2 y'_{t-2} + \cdots + \alpha_p y'_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \tag{3}$$

The modelling process uses the `statsmodels` and `pmdarima` libraries in Python. The dataset is split into training (70%), validation (20%), and test (10%) sets. Stationarity is verified using the Augmented Dickey–Fuller test, which rejects the null hypothesis of a unit root at the 1% significance level.

The optimal ARIMA order of (1,0,2) is chosen using the `auto_arima` function by minimising the Akaike Information Criterion (AIC). The model is first trained on the training set and then evaluated on the validation and test sets using mean absolute error (MAE) and root mean squared error (RMSE).

4.5 Gated Recurrent Unit Network

Gated recurrent unit (GRU) network (Chung et al., 2014) is a type of recurrent neural network (RNN) designed to capture temporal dependencies in sequential data. Basic RNNs suffer from the vanishing gradient problem, which limits their ability to learn long-term patterns. GRUs address this issue through gating mechanisms.

The GRU architecture consists of two gates: the **update gate** (z_t), which determines how much past information is carried forward, and the **reset gate** (r_t), which determines how much past information to forget.

The GRU cell operates according to the following equations:

$$z_t = \sigma\left(W^{(z)} x_t + U^{(z)} h_{t-1}\right) \tag{4}$$

$$r_t = \sigma\left(W^{(r)} x_t + U^{(r)} h_{t-1}\right) \tag{5}$$

$$\tilde{h}_t = \tanh\left(W x_t + r_t \odot U h_{t-1}\right) \tag{6}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{7}$$

where x_t denotes the input at time t, h_t is the hidden state, and $\odot$ denotes element-wise multiplication. $W^{(z)}$, $W^{(r)}$, W and $U^{(z)}$, $U^{(r)}$, U are weight matrices, and σ is the sigmoid function.

Both GRUs and LSTMs can effectively capture long-term dependencies. However, GRUs are more computationally efficient, as they use only two gates (update and reset), compared to the three gates in an LSTM cell (input, forget, output). Moreover, studies such as Goutte et al. (2023) and Niu and Xu (2020) have found that LSTMs may overfit when applied to noisy financial time series data. Given that this study uses high-frequency data, the GRU architecture is considered the more suitable option.

To test the EMH, two GRU models are trained. Each model uses the past 60 minutes of data to forecast the next return. The first model is trained using only historical OHLC prices, trading volumes, and technical indicators to evaluate the weak form EMH. The second model includes all available features, adding macroeconomic indicators (S&P500, Nasdaq100, CPI, Gold, Oil) and sentiment data (Fear & Greed Index) to assess the semi-strong form EMH.

The hyperparameters for both models are optimised using the `Optuna` library:

- **GRU_1:** 256 neurons, dropout rate 0.214, learning rate 0.0024, batch size 256
- **GRU_2:** 128 neurons, dropout rate 0.429, learning rate 0.0087, batch size 512

4.6 Variational Mode Decomposition

Variational mode decomposition (VMD) (Dragomiretskiy and Zosso, 2014) is an adaptive, non-recursive signal processing technique that decomposes a time series into a predefined number (K) of band-limited intrinsic mode functions (BLIMFs). VMD was developed to overcome the limitations in empirical mode decomposition (EMD) (Huang et al., 1998) such as sensitivity to noise (Liu and Chen, 2019). In this study, VMD preprocesses Bitcoin return data to enhance forecast accuracy in the subsequent VMD-GRU-Attention model by reducing high-frequency noise.

The VMD decomposes the original signal $x(t)$ into K modes $\{u_k(t)\}$ with centre frequencies $\{\omega_k\}$ by solving the constrained variational optimisation problem (Dragomiretskiy and Zosso, 2014; Zheng et al., 2021):

$$\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \quad \text{s.t.} \quad \sum_{k=1}^{K} u_k(t) = x(t) \tag{8}$$

Where:

- $\{u_k\} = \{u_1, u_2, \ldots, u_K\}$: Decomposed BLIMF modes
- $\{\omega_k\} = \{\omega_1, \omega_2, \ldots, \omega_K\}$: Corresponding centre frequencies
- $\delta(t)$: Dirac function
- $*$: Convolution operator
- $\left(\delta(t) + \frac{j}{\pi t}\right)$: Hilbert transform to construct analytic signal
- $e^{-j\omega_k t}$: Frequency shifting to baseband
- $\| \cdot \|_2$: L^2 norm measuring the mode's bandwidth

The optimisation via the alternate direction method of multipliers (ADMM) minimises the total bandwidth of all BLIMFs whilst enforcing that their sum reconstructs the original signal. In this study, computational constraints limit decomposition to a maximum of $K = 8$ components.

4.7 Attention Mechanism

The attention mechanism (Luong et al., 2015) allows neural networks to focus on the most relevant parts of an input sequence when generating an output, rather than treating all inputs equally. In time series forecasting, this means assigning higher weights to time steps that are more informative for predicting the next value. This improves the model's ability to capture long-term dependencies and non-linear patterns, especially in high-frequency financial data.

In this study, a self-attention mechanism is applied to the GRU layer's hidden states to identify significant temporal patterns in high-frequency returns. Given an input sequence $\{h_1, ..., h_T\}$ from the GRU, attention weights $a_{t,s}$ are computed using dot-product scoring via the `tf.keras.layers.Attention` function, as shown in Equation 9. These weights are then normalised using the softmax function, as shown in Equation 10:

$$\text{score}(h_t, h_s) = h_t^\top h_s \tag{9}$$

$$a_{t,s} = \frac{\exp(\text{score}(h_t, h_s))}{\sum_{s'=1}^{T} \exp(\text{score}(h_t, h_{s'}))} \tag{10}$$

where h_t and h_s are hidden states from the GRU layer at different time steps, and $a_{t,s}$ denotes the attention weight assigned to h_s when computing the context vector for h_t.

4.8 VMD-GRU-Attention Network

The VMD-GRU-Attention model, inspired by the VMD-AttGRU architecture proposed by (Niu and Xu, 2020), is used to test the weak form EMH by forecasting Bitcoin returns. In this model, VMD is first applied to decompose the noisy return series into K subseries, denoted as BLIMF_1, BLIMF_2, ..., BLIMF_K. Each BLIMF is then used as input to train a separate GRU-Attention model.

The attention mechanism is applied to enable each model to focus on the most relevant parts of the series by assigning different weights to the GRU outputs at each time step. Placing the attention layer after the GRU allows the model to reweight features extracted by the GRU, rather than applying attention to the raw input.

Hyperparameters for each GRU-Attention model are fine-tuned using the `Optuna` library, resulting in 64 neurons, a dropout rate of 0.106, a learning rate of 0.0005, and a batch size of 256. The final prediction is obtained by aggregating the outputs from all GRU-Attention models. Figure 2 shows the full structure of the VMD-GRU-Attention model.

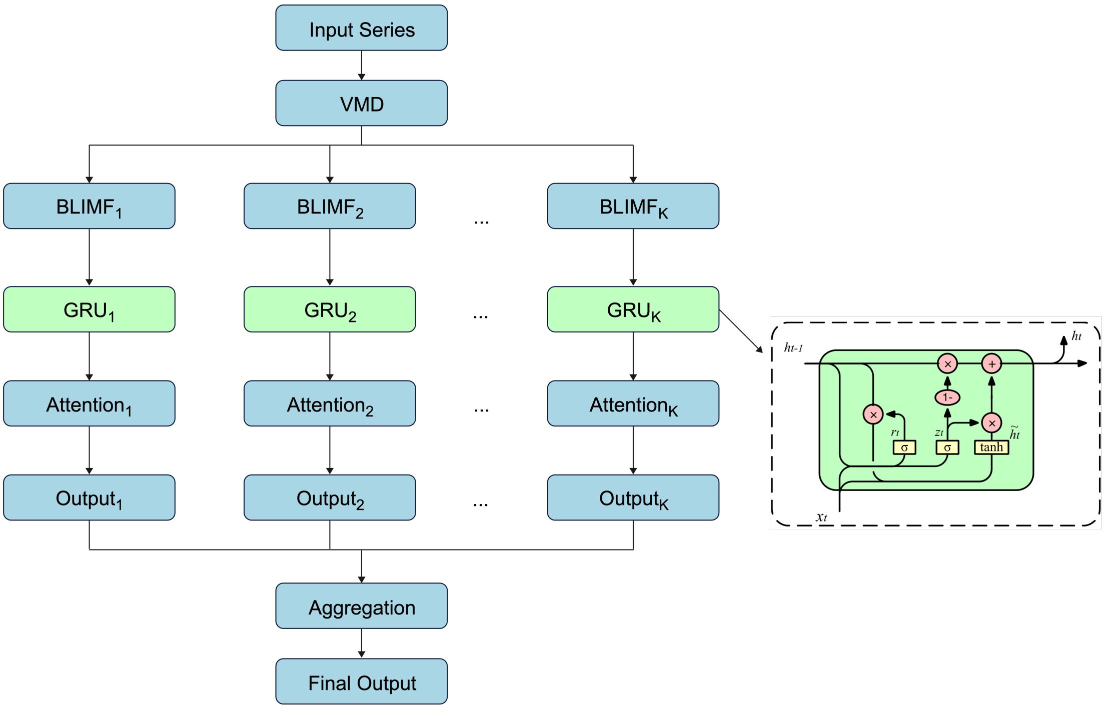


Figure 2: The structure of the VMD-GRU-Attention model.

5 Results

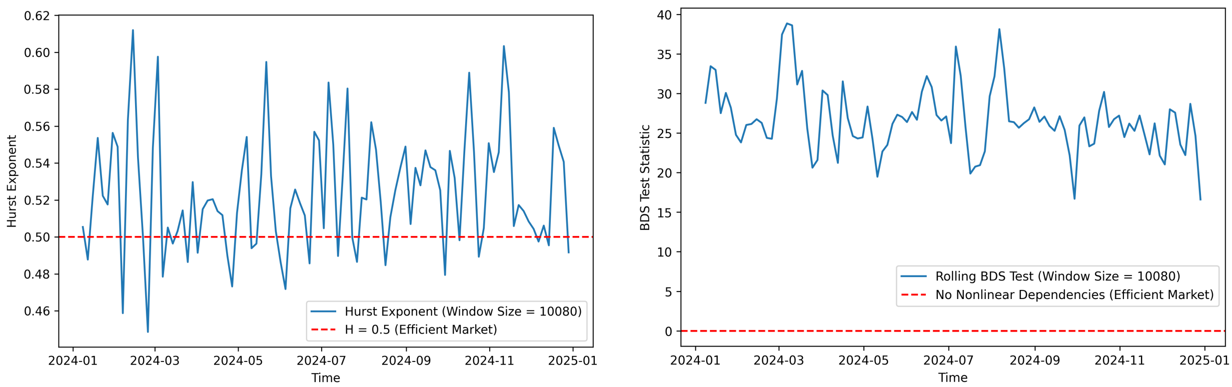
Table 2 presents the results of the statistical tests applied to Bitcoin 1-minute returns. The **Augmented Dickey–Fuller (ADF)** test rejects the null hypothesis of a unit root at the 1% significance level, indicating that the return series is stationary. The **Ljung–Box test** statistic reveals significant autocorrelation, suggesting that past returns contain information that can help predict future returns, thereby violating the weak form EMH. The **Wald–Wolfowitz Runs test** also rejects the null hypothesis of randomness. Similarly, the heteroscedasticity-robust **Variance Ratio test** rejects the random walk hypothesis, and the significantly negative test statistic indicates mean-reverting behaviour in Bitcoin returns (De Nicola, 2021; Yi et al., 2023). The estimated **Hurst exponent** ($H = 0.53$), though close to 0.5, suggests mild persistence or trending in the data (Urquhart, 2016).

However, since prior studies have shown that market efficiency evolves over time (Lo, 2004; Noda, 2021; Yi et al., 2023), a rolling-window approach is implemented. Using a 7-day rolling window (10,080 observations), Figure 3a shows that the Hurst exponent fluctuates between approximately 0.46 and 0.62, with values predominantly above 0.5. This implies a persistent or trending behaviour, where positive or negative returns are likely to be followed by returns of the same sign (Aslam et al., 2023). Finally, the **rolling BDS test** shown in Figure 3b indicates the presence of serial dependence during the whole sample period, further challenging the weak form EMH in the Bitcoin market.

Table 2: Test results for weak form efficiency of Bitcoin returns.

N	ADF test	Ljung–Box test	Runs test	Variance Ratio test	Hurst exponent
509,363	-72.37***	159.10***	-6.10***	-46.28***	0.53

Notes: N denotes the number of observations. ADF refers to the Augmented Dickey–Fuller test. *** indicates significance at the 1% level.



(a) Rolling Hurst exponent.

(b) Rolling BDS test.

Figure 3: Rolling-window analyses of Bitcoin returns.

To further assess the efficiency of the Bitcoin market, a range of econometric and deep learning models, including an ARIMA model, two GRU networks and the proposed VMD-GRU-Attention model are employed to predict minute-level returns. According to the weak form EMH, past information cannot be used to predict future returns. Therefore, if any model consistently outperforms the naïve benchmark, it may suggest a violation of the EMH. Forecast accuracy is primarily evaluated using mean absolute error (MAE) and root mean squared error (RMSE) on the test sets.

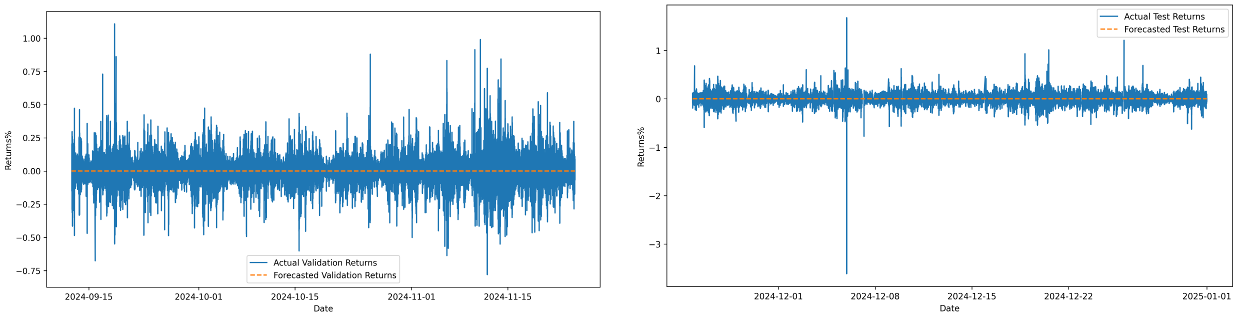
Table 3 shows the coefficients from the ARIMA(1,0,2) model. The first-order autoregressive (AR) coefficient is -0.5057, which suggests that the previous period's return has a negative impact on the current return. The first-order moving average (MA) coefficient (0.4856) is positive, which means that recent shocks continue to influence the current return in the same direction, creating short-term momentum or persistent behaviour. The second-order MA coefficient (-0.0140) is small and negative, meaning that this persistent effect partially reverses after two periods. All coefficients are statistically significant at the 1% level.

However, Figures 4a and 4b show that the model fails to capture the magnitude and direction of most short-term fluctuations. The forecasts remain flat, particularly when compared to the high volatility exhibited in the actual returns. This shows that the ARIMA model has limited forecast power in the highly volatile Bitcoin market.

Table 3: Estimated coefficients from the ARIMA(1,0,2) model.

Term	Coef.	Std. Err.	z	P$>\|z\|$	[0.025	0.975]
ar.L1	-0.5057***	0.112	-4.500	0.000	-0.726	-0.285
ma.L1	0.4856***	0.112	4.322	0.000	0.265	0.706
ma.L2	-0.0140***	0.002	-6.615	0.000	-0.018	-0.010
sigma2	0.0062***	2.79e-06	2207.633	0.000	0.006	0.006

Notes: *** indicates significance at the 1% level. All coefficients are statistically significant with $p < 0.001$.

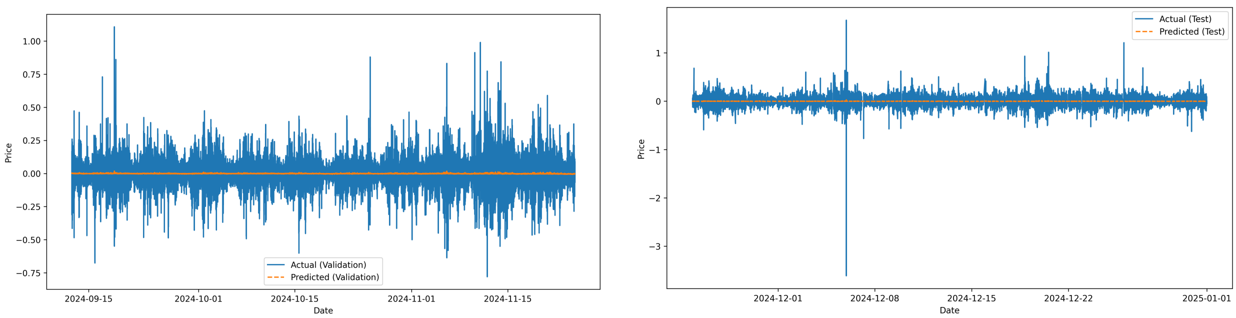


(a) Validation set. (b) Test set.

Figure 4: ARIMA forecasts vs actual returns on validation and test sets.

Figures 5a and 5b compare the forecasted and actual returns on the validation and test sets. The GRU_1 model is trained solely on past prices, trading volumes, and technical indicators to assess whether the weak form EMH holds in the Bitcoin market. Similar to the ARIMA model, the predicted returns remain relatively flat and fail to capture the large spikes and dips in the actual returns across both sets. Moreover, the naïve model outperforms GRU_1 in terms of MAE on both the validation and test sets. This suggests that the Bitcoin market may exhibit weak form efficiency.

To explore whether including more market information can improve predictive performance, the GRU_2 model is trained using additional macroeconomic and sentiment data to assess the semi-strong form of market efficiency. However, this model only marginally improves the MAE and still underperforms compared to the naïve model. Its forecasts also remain relatively flat[1], failing to capture short-term fluctuations, as is the case with GRU_1. These results provide no strong evidence against the semi-strong form EMH, suggesting that publicly available macroeconomic and sentiment indicators may not offer much advantage in predicting Bitcoin returns. Whilst the Bitcoin market may deviate from perfect efficiency, exploiting such inefficiencies using historical and public data remains highly challenging.



(a) Validation set. (b) Test set.

Figure 5: GRU_1 forecasts vs actual returns on validation and test sets.

To better capture the high volatility in 1-minute Bitcoin returns, the more sophisticated VMD-GRU-Attention model is employed. This hybrid model combines variational mode decomposition (VMD), GRU, and an attention mechanism to detect complex, non-linear hidden patterns in high-frequency data. VMD

[1]See Figures 9a and 9b in the appendix.

decomposes the noisy return series into band-limited intrinsic mode functions (BLIMFs), as shown in Figure 6. This helps isolate frequency-specific behaviours and reduce the negative impact of high-frequency noise (Niu and Xu, 2020). Each BLIMF is then used to train a separate GRU-Attention model. The attention mechanism, applied after the GRU, allows the network to focus on the most informative parts of the sequence.

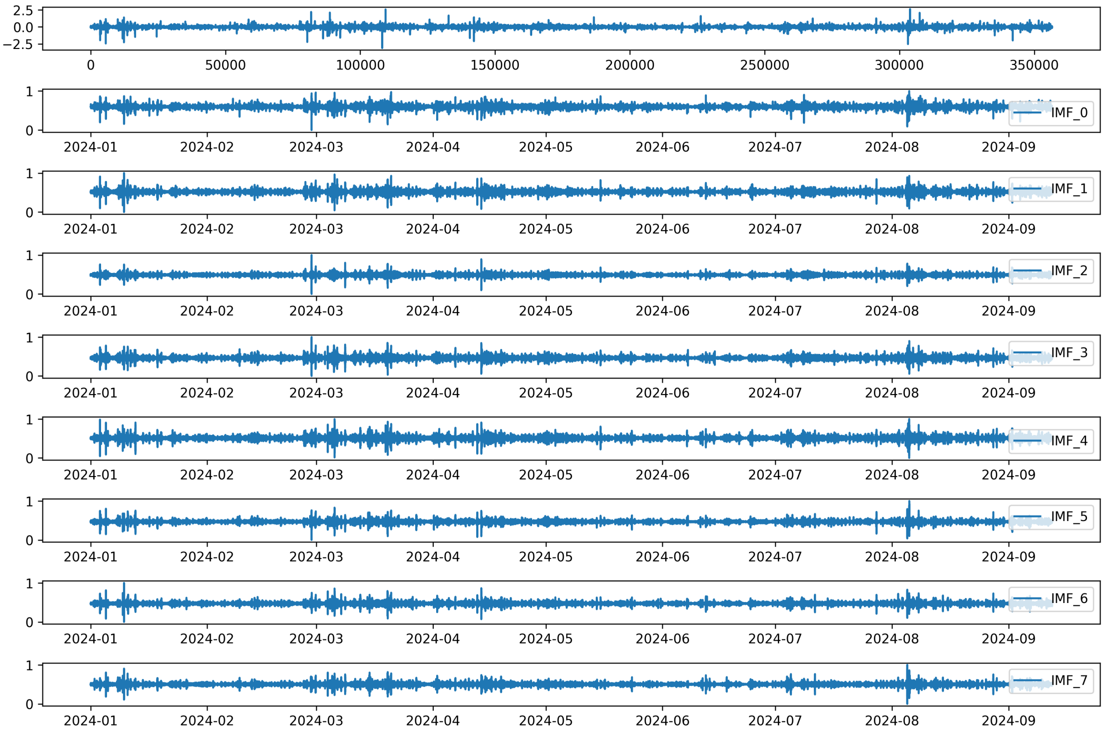
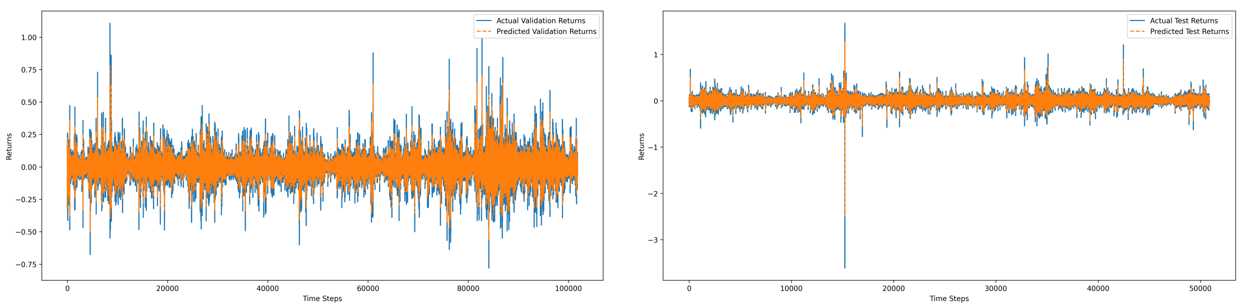


Figure 6: VMD decomposition of Bitcoin returns in the training set. Note that BLIMFs are labelled as IMFs in this figure.

As shown in Figures 7a and 7b, the VMD-GRU-Attention model captures fluctuations in Bitcoin returns far more accurately than previous models with a substantial improvement in both MAE and RMSE, as shown in Table 4. These results challenge the weak form EMH and are consistent with findings in the existing literature.



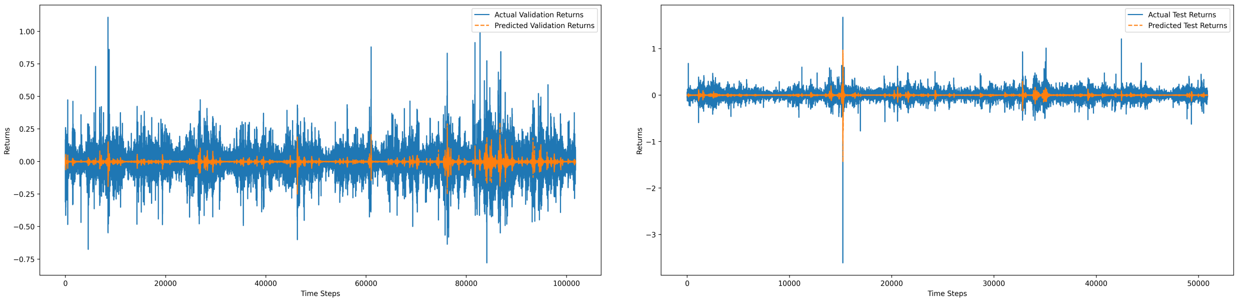
(a) Validation set.

(b) Test set.

Figure 7: VMD-GRU-Attention forecasts vs actual returns on validation and test sets.

To assess whether reversing the order of the attention and GRU layers in the VMD-AttGRU model from Niu and Xu (2020) can actually improve the model's performance, the original VMD-AttGRU architecture is reimplemented and fine-tuned. However, it underperforms considerably compared to the proposed VMD-GRU-Attention model (see Table 4). Although the VMD-AttGRU model captures some fluctuations in returns (see Figures 8a and 8b), it is less effective than the proposed architecture. Therefore, placing the attention layer after the GRU layer, as in the proposed model, results in better predictive performance.

Table 4 summarises the forecast performance across all models. The ARIMA model produces identical MAE and RMSE to the naïve benchmark, as it effectively collapses to predicting the mean return, likely due



(a) Validation set.

(b) Test set.

Figure 8: VMD-AttGRU forecasts vs actual returns on validation and test sets.

to the extreme volatility of Bitcoin. The VMD-GRU-Attention model achieves the lowest forecast errors, whilst the GRUs unexpectedly underperform both the ARIMA and naïve models.

Together with the statistical tests, the strong performance of the VMD-GRU-Attention model provides further evidence against the weak form EMH in the Bitcoin market. Nevertheless, consistently exploiting profitable trading opportunities remains a challenge, given the market's highly volatile nature.

Table 4: Forecast performance of models on validation and test sets (70-20-10 split).

Model	Val. MAE	Val. RMSE	Test MAE	Test RMSE
Naïve	0.0412	0.0651	0.0476	0.0749
ARIMA	0.0412	0.0651	0.0476	0.0749
GRU_1	0.0414	0.0651	0.0480	0.0749
GRU_2	0.0413	0.0651	0.0478	0.0749
VMD-AttGRU	0.0398	0.0615	0.0457	0.0703
VMD-GRU-Attention	0.0141	0.0211	0.0158	0.0239

Note: MAE and RMSE denote mean absolute error and root mean squared error.

5.1 Robustness Checks

This section examines the robustness of the models by (i) applying different train-validation-test splitting ratios (60-20-20 and 50-30-20), and (ii) removing half of the training set. Table 5 presents the results. The findings show that changes in the splitting ratio do not affect the relative performance of the models. Forecast errors (MAE and RMSE) remain consistent with those observed in the original split. However, when the training set is halved, prediction errors for both basic GRU models increase drastically, possibly due to overfitting. In all cases, the VMD-GRU-Attention model continues to outperform the others, achieving the lowest errors by a wide margin. Its consistently strong performance across different sample sizes demonstrates its robustness.

Table 5: Forecast performance (Test MAE and RMSE) of models across robustness checks.

Model	60-20-20		50-30-20		Reduced Training	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
Naïve	0.0468	0.0731	0.0468	0.0731	0.0476	0.0749
ARIMA	0.0468	0.0731	0.0468	0.0731	0.0476	0.0749
GRU_1	0.0471	0.0731	0.0541	0.0778	0.2972	0.3078
GRU_2	0.0470	0.0731	0.0534	0.0774	0.3008	0.3113
VMD-GRU-Attention	0.0169	0.0253	0.0181	0.0270	0.0274	0.0405

Notes: The VMD-GRU-Attention model consistently achieves the lowest errors. GRU models show a marked decline in performance when trained on reduced data.

6 Conclusion

This study investigates whether the Bitcoin market in 2024 follows the weak form efficient market hypothesis (EMH) using a combination of traditional methods and modern deep learning models. Minute-level data from 2024 are used to compute log returns and technical indicators. A variety of statistical tests, including the Ljung-Box test, Runs test, Variance Ratio test, Hurst exponent, and rolling BDS test, are applied to examine the presence of memory in historical Bitcoin returns. In addition, ARIMA, GRU, and the VMD-GRU-Attention models are used to forecast Bitcoin returns to further assess the weak form EMH.

Results overwhelmingly suggest that the 2024 Bitcoin market does not follow the weak form EMH. The statistical tests detect significant autocorrelation in the return series, consistent with existing literature. Although the ARIMA and the two basic GRU models fail to outperform the benchmark in forecasting out-of-sample returns, the proposed VMD-GRU-Attention model achieves strong performance across both evaluation metrics (MAE and RMSE) on the validation and test sets.

These findings suggest that profitable trading opportunities remain in the Bitcoin market, although exploiting them likely requires sophisticated algorithms. The results also imply that further regulation for the Bitcoin market may be necessary to build a safer environment and increase confidence among both retail and institutional investors.

However, this study faces a few limitations. Computational constraints restrict the number of band-limited intrinsic mode functions (BLIMFs) that can be extracted through the VMD. A deeper decomposition may enhance forecast performance, but hardware limitations prevent this. Similarly, it is not feasible to collect large volumes of sentiment data directly from social media platforms, such as Twitter or Reddit. As a result, the study relies on the daily Crypto Fear & Greed Index as a proxy for market sentiment, which may reduce the model's performance, given that minute-level sentiment would better align with the 1-minute return series.

Future work may expand on this analysis by including other major cryptocurrencies, such as Ethereum, Tether, and Solana, to evaluate whether they exhibit similar levels of market efficiency and to develop trading strategies that test whether consistent profits can still be made after factoring in trading costs.

References

Aslam, F., Memon, B. A., Hunjra, A. I., and Bouri, E. (2023). The dynamics of market efficiency of major cryptocurrencies. *Global Finance Journal*, 58:100899.

Aslan, A. and Sensoy, A. (2020). Intraday efficiency-frequency nexus in the cryptocurrency markets. *Finance Research Letters*, 35:101298.

Bariviera, A. F. (2017). The inefficiency of bitcoin revisited: A dynamic approach. *Economics Letters*, 161:1–4.

Bouteska, A., Sharif, T., Isskandarani, L., and Abedin, M. Z. (2025). Market efficiency and its determinants: Macro-level dynamics and micro-level characteristics of cryptocurrencies. *International Review of Economics and Finance*, 98:103938.

Box, G. E. and Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Broock, W. A., Scheinkman, J. A., Dechert, W. D., and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, 15(3):197–235.

Chen, Y. C. (2021). Empirical analysis of bitcoin price. *Journal of Economics and Finance*, 45(4):692–715.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint*.

De Nicola, G. (2021). On the intraday behavior of bitcoin. *Ledger*, 6:40–54.

Dragomiretskiy, K. and Zosso, D. (2014). Variational mode decomposition. *IEEE Transactions on Signal Processing*, 62(3):531–544.

Fama, E. F. (1970). Efficient capital markets. *The Journal of Finance*, 25(2):383–417.

Fernandes, L. H., Bouri, E., Silva, J. W., Bejan, L., and de Araujo, F. H. (2022). The resilience of cryptocurrency market efficiency to covid-19 shock. *Physica A: Statistical Mechanics and its Applications*, 607:128218.

Goutte, S., Le, H.-V., Liu, F., and Von Mettenheim, H.-J. (2023). Deep learning and technical analysis in cryptocurrency market. *Finance Research Letters*, 54:103809.

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

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995.

Huang, Q. (2024). Bitcoin price prediction based on fear & greed index. *SHS Web of Conferences*, 181:02015.

Huang, Z.-C., Sangiorgi, I., and Urquhart, A. (2024). Forecasting bitcoin volatility using machine learning techniques. *Journal of International Financial Markets, Institutions and Money*, 97:102064.

Jaquart, P., Dann, D., and Weinhardt, C. (2021). Short-term bitcoin market prediction via machine learning. *The Journal of Finance and Data Science*, 7:45–66.

Jaquart, P., Köpke, S., and Weinhardt, C. (2022). Machine learning for cryptocurrency market prediction and trading. *The Journal of Finance and Data Science*, 8:331–352.

Kakinaka, S. and Umeno, K. (2022). Cryptocurrency market efficiency in short-and long-term horizons during covid-19: An asymmetric multifractal analysis approach. *Finance Research Letters*, 46:102319.

Kang, H.-J., Lee, S.-G., and Park, S.-Y. (2022). Information efficiency in the cryptocurrency market: The efficient-market hypothesis. *Journal of Computer Information Systems*, 62(3):622–631.

Kotu, V. and Deshpande, B. (2019). Chapter 12 - time series forecasting. In Kotu, V. and Deshpande, B., editors, *Data Science (Second Edition)*, pages 395–445. Morgan Kaufmann, second edition edition.

Liu, H. and Chen, C. (2019). Data processing strategies in wind energy forecasting models and applications: A comprehensive review. *Applied Energy*, 249:392–408.

Ljung, G. M. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Lo, A. W. (2004). The adaptive markets hypothesis: Market efficiency from an evolutionary perspective. *The Journal of Portfolio Management*, 30(5):15–29.

Lo, A. W. and MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies*, 1(1):41–66.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint*.

Nadarajah, S. and Chu, J. (2017). On the inefficiency of bitcoin. *Economics Letters*, 150:6–9.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. White paper.

Niu, H. and Xu, K. (2020). A hybrid model combining variational mode decomposition and an attention-GRU network for stock price index forecasting. *Mathematical Biosciences and Engineering*, 17(6):7151–7166.

Noda, A. (2021). On the evolution of cryptocurrency market efficiency. *Applied Economics Letters*, 28(6):433–439.

Pourrezaee, A. and Hajizadeh, E. (2024). Forecasting bitcoin volatility and value-at-risk using stacking machine learning models with intraday data. *Computational Economics*.

Urquhart, A. (2016). The inefficiency of bitcoin. *Economics Letters*, 148:80–82.

Wald, A. and Wolfowitz, J. (1940). On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162.

Yi, E., Yang, B., Jeong, M., Sohn, S., and Ahn, K. (2023). Market efficiency of cryptocurrency: evidence from the bitcoin market. *Scientific Reports*, 13(1):4789.

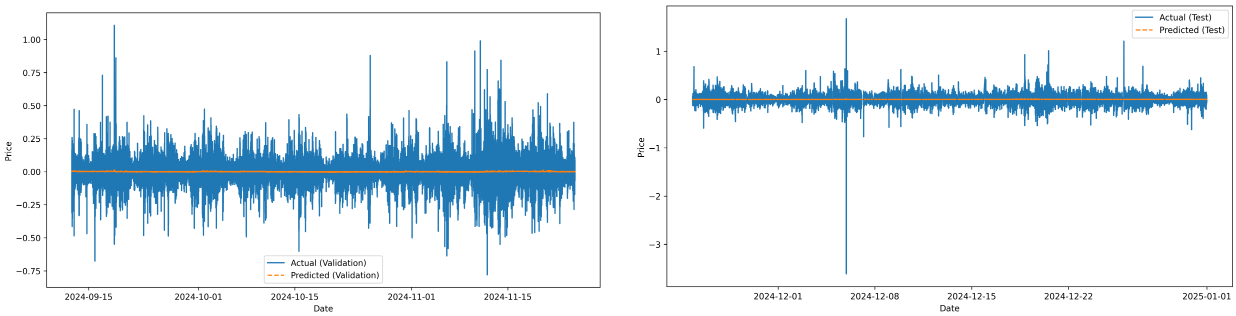
Zheng, D., Zhang, W., Alemu, S., Wang, P., Bitew, G., Wei, D., and Yue, J. (2021). Chapter 4 - short-term renewable generation and load forecasting in microgrids. In *Microgrid Protection and Control*, pages 57–96. Academic Press.

Appendix

Table 6: Description of variables from the main dataset *Bitcoin_data_2024_w_TI_n_Macro.csv*.

Variable	**Description**
date	Timestamp of the observation
Open	Open price of Bitcoin
High	Highest price within the time interval
Low	Lowest price within the time interval
Close	Close price of Bitcoin
VolumeBTC	Trading volume of Bitcoin (in BTC)
Returns%	Percentage log returns
SMA_20	20-period simple moving average
Volatility_20	20-period rolling standard deviation
RSI	Relative strength index
MACD_Hist	Moving average convergence divergence
ROC	Rate of change
CCI	Commodity channel index
WilliamsR	Williams %R
ADX	Average directional index
SP500	S&P 500 composite index
CPI	US consumer price index
Gold	Gold bullion price (New York, USD/ounce)
Nasdaq100	Nasdaq-100 index
VIX	CBOE volatility index (VIX)
Oil	Brent crude oil price (USD/barrel)
fear_n_greed_index	Crypto fear & greed index (sentiment measure)

Notes: This table summarises the variables used in the statistical analysis and modelling of Bitcoin returns using 1-minute interval data. The dataset includes technical indicators, macroeconomic variables, and sentiment-based measures.



(a) Validation set.

(b) Test set.

Figure 9: GRU_2 forecasts vs actual returns on validation and test sets.