Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

# Multimodal Transformer with Attention for High-Frequency Bitcoin Price Forecasting: Integrating Market, Sentiment, and Macroeconomic Signals

**Muhammad Ghazanfar Ullah Khan**
Department of Computer Systems Engineering, UIT University Karachi
Email: ghazanfar.ullah@gmail.com

**Khaliq Ahmed**
Department of Computer Science, Nazeer Hussain University, Karachi
Email: drkhaliq.ahmed@nhu.edu.pk

**Syed Sheraz Ul Hasan Mohani**
Department of Electrical Engineering, Iqra University, Karachi
Email: smohani@iqra.edu.pk

**Maheen Danish**
Computer Engineering Department, Sir Syed University of Engineering and Technology, Karachi Email: msadiq@ssuet.edu.pk

**Simra Nasir**
NED University of Engineering and Technology, Karachi
Email: simmra244@gmail.com

**Abdul Khaliq**
Department of Computer Science, Institute of Business Management, Karachi
Email: khaliq@iobm.edu.pk

## ABSTRACT

Forecasting Bitcoin prices for the near term is not easy because of the crypto market's rapid and complex nature. Therefore, I have created a Transformer model that can predict hourly Bitcoin prices with the help of attention. It gets into the price data as well as market, sentiment, and VIX (Volatility Index) factors. A comparison with classical methods such as ARIMA, XGBoost, and LSTM was performed in order to assess the model's performance. The outcome indicates that Transformer model outperformed all the others with a MAE of $3.26 and a RMSE of $4.57. Moreover, the model's attention mechanism reveals the factors it is most focused on, which allows better understanding of the predictions it makes. Thus, it seems that Transformers are "wizards" in this kind of rapid market guessing, offering not only accurate forecasts but also some insights into the underlying driver(s) of the predictions, which turns out to be beneficial for decision-making.

**Keywords:** ARIMA, Bitcoin, LSTM, XGBoost

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

## Introduction

The growing trend of Bitcoin, a non-physical and decentralized currency, has a significant effect on the global monetary system [1]. Bitcoin is a currency that can be traded directly between two users, without any intermediary, it uses a distributed ledger-based technology (blockchain) and all its transactions are visible and require no trust. Over the 10 years period, Bitcoin has gained enormous acceptance not only among the investors and traders but also among researchers—a group which is probably the least expected to accept it—among others [2]. Nevertheless, the most concerning characteristic of Bitcoin still remains its price movement, which is frequently referred to as highly volatile and unmanageable; this is mainly a reason for his lesser acceptance in most areas. One of the determinants that might contribute to the insecurity surrounding the future price of Bitcoin is the sophisticated and ever-changing relationship of numerous factors like trading volume, demand, regulations, financial market trends, media attention, etc. [3][4]. A highly precise prediction of Bitcoin prices is of paramount importance for performing trade with knowledge, planning investments and managing risks. Financial time series forecasting has made use of statistical models such as Auto-Regressive Integrated Moving Average (ARIMA) [5]. These models have some merits, but they are not generally able to bring the characteristics of price data, which are at the same time non-linear, non-stationary, and highly stochastic for cryptocurrencies, into their framework and therefore are not applicable in such situations [6].

One of the most important challenges in this research area is the development of accurate future Bitcoin price prediction models that are reliable and incorporate the signals, price changes, and the impact of outside factors [7]. Machine and deep learning techniques, which have been used to get quite a good accuracy on standard datasets, still have to deal with the consistency issue across different time series, the integration of various factors such as economic indicators and market sentiment, and the adaptation to real-time market changes [8]. The financial sector is consequently looking for forecasting systems that will be accurate, universal, and quick to respond thus making more reliable and practical decisions possible. Moreover, the everchanging and unpredicted nature of the cryptocurrency market compels the models to be updated in real time to include not only the new patterns that they will have to predict but also the old ones that they will have to re-predict [9][10]. A more reliable, widely estimated, and adaptive architecture is needed to support the existing predictive tools in the financial market with accuracy and practical applications. The research concerning hourly prediction of Bitcoin price eloquently illustrates the performance of traditional (ARIMA), machine learning (XGBoost, LSTM), and a new proposed Multimodal Transformer with Attention. The superiority of the new method is reflected in the regard to both accuracy and interpretability.

## LITERATURE REVIEW

The increasing use of cryptocurrency, particularly Bitcoin, has greatly inspired researchers to investigate the decentralization of Bitcoin, its market behavior, and the use of machine learning in predicting its price. Several research works have dealt with the matter regarding the unique primary structure of Bitcoin as a currency in the digital world and were testing its market trends using various statistical and machine

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

learning methods. Meenu Gupta and collaborators describe Bitcoin as a decentralized and peer-to-peer currency, and they also recognize the role of "miners" who manage the transactions and put new Bitcoin into circulation, which is capped at a total of 21 million units [11]. By statement of January 30, 2021, there were around 18,614,806 Bitcoins that had been mined and were already trading in the market. This study utilizes a dataset from Kaggle with characteristics such as Date, High, Low, Open, Close, Volume, and Market Cap. The research methodology consisted of data cleansing, conducting Exploratory Data Analysis (EDA) to reveal trends, and testing the effectiveness of various machine learning algorithms such as Ordinary Least Squares, Logistic Regression, IBK, Random Subspace, Stacking, and Input Mapped Classifiers. Jyoti Verma and her colleagues introduced a hybrid deep neural network model which consists of both CNN and LSTM models for univariate time series forecasting [12]. The proposed model built upon the benefits of feature extraction from CNNs and the ability of LSTM to deal with long-term sequences. The model was tested on two real-world datasets: electricity consumption and air quality. The model's effectiveness was evaluated against four other recurrent neural network models: RNN, LSTM, GRU, and Bidirectional LSTM. Throughout the experiments, it was found that not only did the hybrid CNN-LSTM model attain the highest performance but it was also very durable against noise, and the presence of convolutional layers was the main reason for its better prediction accuracy. In their paper, Dhawani et al. provided an exhaustive review accompanied by a classification of existing time series forecasting methods that covered traditional econometric models along with state-of-the-art machine learning techniques [13]. The application of time series techniques is gradually becoming more common and thereby used in the rendering of accurate models, being very difficult to achieve because of the constantly changing, complex, and chaotic nature of the data. The author divides the methods into two categories: traditional statistical models (linear models such as AR, MA, ARMA, ARIMA and smoothing techniques like SES, DES, TES; and nonlinear algorithms like ARCH and GARCH) and machine learning models where he emphasizes their contrasting characteristics. He indicates that the main issues in time series predicting are the often-unavoidable data preprocessing due to poor data quality problems such as noise and missing values, and the limitation of real-time model updating in big-data environment to maintain forecast precision which often requires rolling window techniques. It is observed that classical models do give accurate forecasts but only in situations where the volume is very high.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

## Methodology

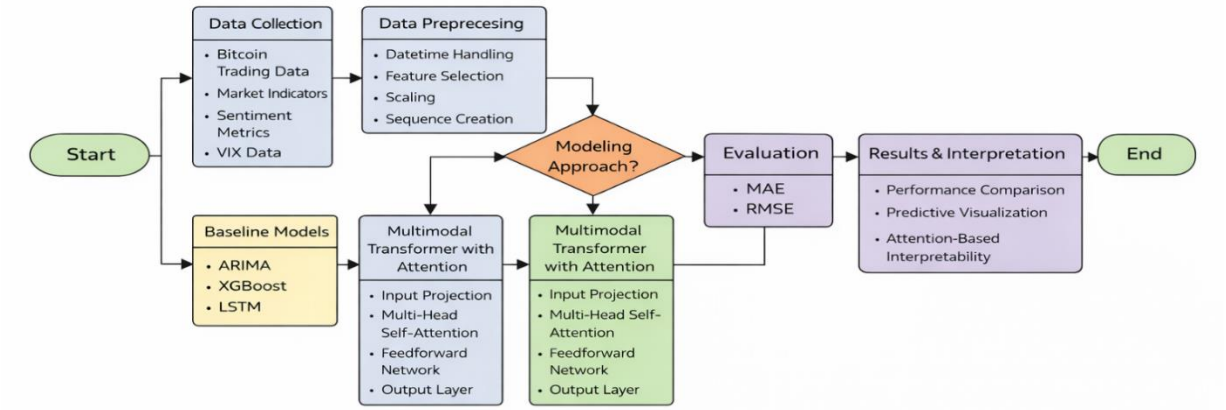



Figure 1: overall methodology

The diagram 1 reveals a thorough workflow of Bitcoin price prediction using traditional and deep learning models. The operation starts by collecting data from various places, which consist of Bitcoin trading data, market indicators, sentiment metrics, and VIX data. Then, after preprocessing steps such as handling of datetime, selection of features, normalization, and sequence creation, dividing of dataset into training and testing sets occurs. There is a decision point to differentiate between baseline models (ARIMA, XGBoost, and LSTM) and the new multimodal Transformer with attention in the modeling routes. Both modeling paths intersect at the evaluation stage where MAE and RMSE are the metrics used, followed by interpretation of results, comparison of performance, and application of attention-based explainability.

### Dataset Description

The authors of the investigation rely on the two-year period of hourly Bitcoin (BTC/USDT) market data comprising 17,515 observations in making short-term cryptocurrency price predictions. The data is sorted as a multimodal time series that chronicles the events that cause Bitcoin price fluctuations and hence, the one that covers all aspects of price dynamics, trading activity, market structure indicators, investor sentiment, and macroeconomic risk measurements has been formed. The hourly data is the main source of the intraday price behavior, volatility, and trend creation, and it includes Open, High, Low, and Close (OHLC) prices, as well as trading volume, which is considered the main factor affecting market liquidity and trading participants. These variables form the foundation of the short-term price prediction as they reflect the changes in supply-demand and short-term momentum

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

most closely. To cover market microstructure and derivatives market behavior, the dataset has been created by combining perpetual futures market funding rates and their open interest. The funding rate provides the whole scenario of positions—longs and shorts—whereas open interest illustrates the total number of outstanding derivative contracts and, consequently, the extent of speculation in the market and possible price shifts. The Fear and Greed Index, which transforms volatility, momentum, and social signals into a single numeric range of 0-100, is applied to determine the investors' mood and the extremes signify fear and greed thereby revealing the psychological factors that influence the market. Public interest and retail sentiment are illustrated by Google Trends search volumes for "buy crypto" and "bitcoin", which act as proxies for market participation driven by attention. Moreover, general financial risk conditions are measured using the Open, High, Low, and Close values of the CBOE Volatility Index (VIX). As a widely acknowledged measure of global market uncertainty, the VIX indicates how macroeconomic and risk-averse conditions might leak into the cryptocurrency markets especially when volatility is high.

The hourly cryptocurrency data, which are very high-frequency, sometimes contain missing values resulting from the unavailability of exchanges or the partial reporting of data. In order to maintain the integrity of the time series and provide uniform inputs to the model, the missing observations were treated through forward-fill imputation, where the last accessible value is carried over. By this method, all sliding window sequences of 24 hours are kept complete for the purposes of model training and evaluation, thus the temporal structure is preserved.

**Summary Statistics**

Table 1

| **Feature Category** | **Examples** | **Count** |
|---|---|---|
| Price & Volume | Open, High, Low, Close, Volume | 5 |
| Market Indicators | Funding Rate, Open Interest, Fear & Greed Index | 3 |
| Sentiment Metrics | Google Trends "buy crypto", "bitcoin" | 2 |
| Macroeconomic Risk | VIX Open, High, Low, Close | 4 |
| **Total Features** | | 14 |

The complete set of data includes 17,515 hourly recordings that were put together in 24-hour sliding window sequences and used as inputs for the predictive models. With this temporal modeling, the models can acquire short-term dependencies while maintaining the sequential structure of high-frequency cryptocurrency data. The dataset is a complete multimodal representation of the factors influencing Bitcoin price movements by combining price and volume information with market indicators, investor sentiment measures, and macroeconomic risk variables. This combination gives the models the ability to capture quick intraday movements, measure the prevailing market sentiment, and incorporate financial risk conditions around them. The proposed framework thus leads to predictions that are both accurate and

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

interpretable, especially during times of increased volatility and rapid changes in market dynamics.

**Data Preprocessing**

High-frequency financial time series data like hourly Bitcoin prices demand a lot of preprocessing effort so that the machine learning and deep learning models can capture the temporal patterns and the multimodal relationships associated with them. The structured preprocessing pipeline established for this purpose first ensured that the temporal integrity was maintained, then numerical stability was enhanced, and finally, realistic model evaluation was promised. First of all, the Datetime column was transformed from string format to Python datetime objects. Then the dataset was arranged in chronological order and timestamp was used for indexing. The correct manipulation of datetime is extremely critical for time series forecasting as it maintains the natural order of the observations and at the same time makes possible the creation of input windows that indeed mirror the flow of time. In addition, this stage also plays a role in enhancing the countermeasures against information leakage by ensuring that no future observations are by mistake included in past predictions.

Subsequently, a targeted feature selection approach was utilized to keep only those variables that are directly linked to Bitcoin price movements. The variables that were selected are the Bitcoin price and volume (OHLCV), the indicators of the derivative market such as funding rate and open interest, the sentiment-based metrics like the Fear and Greed Index and Google Trends search interest, and finally, the macroeconomic risk indicators which are represented by the Open, High, Low, and Close prices of the CBOE Volatility Index (VIX). By limiting the feature set to multimodal signals related to Bitcoin, noise and spurious correlations are reduced while providing the model with economically significant inputs. Missing values in the chosen features were addressed by using forward-fill imputation so as to keep the temporal continuity intact.

In order to handle the different scales of the selected features, all variables were subject to MinMax scaling normalization to the range of [0,1]. This procedure avoids the scenario where features with large numerical ranges like that of Bitcoin prices would overshadow the learning process compared to the bounded sentiment or risk indicators. Besides, feature scaling results in quicker and more stable convergence during training, especially when using gradient-based deep learning models like LSTMs and Transformers.

Subsequently, the normalized data were converted into fixed-length input sequences using a sliding window technique. More precisely, the sliding windows of 24-hours were made which resulted in the input sequences having a shape of (24,14), where each sequence comprised one day of historic data across all the chosen features. In this manner, the Bitcoin closing price for the next hour was linked to every input window facilitating subsequent prediction. The representation based on sequence enables the models to detect the relationships in time and the shifting market dynamics gradually because those are not observable through individual hourly frequency. The final phase of the operation consisted of temporally partitioning the data set into training and testing subsets, assigning 80% of the sequences to training and 20% to testing. This type of partitioning, unlike random splitting, guarantees that

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

no future data is used for the prediction of past values, thus completely eliminating data leakage and very closely imitating the real-life forecasting scenarios.

Nowadays, this preprocessing pipeline supports the factors of chronological assignment, multimodal relevance, numerical stability, and temporal awareness besides making realistic model evaluation easier. The procedures allow for a fair comparison and reliable short-term Bitcoin price prediction by providing a strong foundation for the evaluation of classical statistical methods (ARIMA), machine learning techniques (XGBoost), deep learning baselines (LSTM), and proposed Transformer with Attention architecture.

**Baseline Models**

To see how well our multimodal Transformer with attention works, we tried out three common models: ARIMA, XGBoost, and LSTM. They're like the classic ways to do things in statistics, machine learning, and deep learning. We made sure each one was set up right and tested them all using the same data and rules, so we could compare them fairly.

3.3.1 Autoregressive Integrated Moving Average (ARIMA)

We used an ARIMA model to predict the price of Bitcoin in hourly terms (BTC/USDT close). This is a well-known method for forecasting time series and works well with linear data. Because of the large fluctuations in prices of cryptocurrencies, we transformed the data into a stationary series by using first-order differencing (d = 1). After carrying out the autocorrelation test, I settled on an ARIMA model of (5,1,0). This shows that it is taking into account the last five hours to predict the next hour, although in this case no moving average is applied. We trained the model using the SARIMAX framework, and thereafter the model's performance was measured by applying one-step-ahead predictions. The model's accuracy was determined through calculations of MAE and RMSE using the actual price values. ARIMA offers the advantages of interpretability and computational efficiency, but its restriction to one variable only and linear dynamics has made it less suitable for the volatile and diverse cryptocurrency markets. However, it still serves as a strong statistical reference point.

**Extreme Gradient Boosting (XGBoost)**

For our machine learning experiments, we chose to start with XGBoost (Extreme Gradient Boosting), which is a powerful tool for reasoned analysis and detailed understanding of complicated situations. While ARIMA is limited to time-based analysis with certain restrictions, XGBoost is able to analyze the relationship and the interaction among different kinds of data as well as their simultaneous effect.

XGBoost's ability to handle messy data and discovering the difficult relationships is what made us choose it as the main tool for our project. However, it's not that good at temporal data and does not really capture the order of events unless you explicitly instruct it to do so, thus it might overlook long-term recurring patterns.

This is not the case, however, with XGBoost, which remains a powerful, transparent, and scalable starting point for forecasting that incorporates many factors, such as stock prices.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

**Long Short-Term Memory (LSTM)**

The primary deep learning approach along with a Long Short-Term Memory (LSTM) network was used to analyze how the time series data evolves over time. LSTM networks are well known to solve the gradient problem, which is why they are excellent in detecting long-term trends. After that, the data was sliced into 24-hour pieces, thus the sequences were of shape (24, 14). The architecture of the network included LSTM layers one on top of the other with a fully connected layer at the end for predicting the closing price of Bitcoin for the next hour. The model was trained with the Adam optimizer and the mean squared error (MSE) was used to monitor its performance. LSTM models excel in identifying trends across time periods, but they do not differentiate among features. There is no mechanism for them to indicate the importance of certain features or time steps. Also, LSTMs struggle with large variations in financial data. Moreover, processing of long sequences requires a large amount of computing power.

**Baseline Comparison Summary**

Together, these baseline models provide complementary perspectives:

**ARIMA** captures linear temporal dynamics in univariate price data.

**XGBoost** models nonlinear relationships across multiple features without temporal memory.

**LSTM** learns sequential dependencies but lacks interpretability and explicit feature attention.

The limitations of these approaches motivate the adoption of a **Transformer-based architecture with attention**, which can simultaneously model long-range temporal dependencies, multimodal feature interactions, and provide interpretability through attention weights.

**Multimodal Transformer with Attention**

The suggested approach of a Multimodal Transformer with Attention is purposed to model the intricacy and extreme volatility that characterize hourly Bitcoin price fluctuations by concurrently acquiring knowledge from price data, market indicators, investor sentiment, and macroeconomic risk factors. The input stage consists of 14 multimodal features per time step, which are transformed into a latent representation of higher dimensions of the model size of 64. Such transformation is achieved through the use of a fully connected (Dense) layer that is applied separately to each time step.

The input embedding layer is of great significance in two ways. To begin with, it maps the raw feature space to a dimensionality that suits the Transformer architecture, thus making self-attention operations efficient. The second purpose is that it allows the model to learn complex interactions among different features by embedding heterogeneous inputs like price levels, sentiment indices, and volatility measures into a common representation space. The integration of the embeddings makes it possible to model different sources of information at the same time and provides the Transformer with the ability to detect minute relationships across modalities, which is

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

vital for the precise prediction of Bitcoin prices in the short term.
Mathematically:
$X_{proj} = W_{proj} X + b_{proj}, X \in R^{T \times F}$ , $X_{proj} \in R^{T \times D}$
Where:
(T) = number of time steps (24 hours)
(F) = original feature dimension (14)
(D) = embedding dimension (64)

**Multi-Head Self-Attention**

The multi-head self-attention technique is the heart of the suggested Transformer model that lets the system understand the interplay of time steps and multimodal features while estimating the next-hour Bitcoin closing price. The self-attention mechanism gives the model the ability to continuously weigh and identify past data, thus revealing the most powerful time slices and data sources that change with different market conditions.

Four attention heads are utilized in this research to identify different dependency patterns in the data. A separate representation of the relationships at the temporal and feature level is acquired by each head, which can be, for example, short-term momentum effects, sudden market shocks, sentiment-driven price changes, or wider macroeconomic influences. The simultaneous operation of these attention heads allows for different perspectives on the same input sequence, thus improving the model's potential to discover important patterns in the difficult and unclear financial data.

Multi-head self-attention has an advantage over recurrent neural networks (RNNs) and long short-term memory (LSTM) models in that it can effectively model long-range dependencies without the vanishing gradient problem. Furthermore, joint processing of multimodal features by the model enables it to easily shift its attention to the price movements, investor sentiment, or market indicators depending on the situation, thus increasing both predictive accuracy and interpretability.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q K^T}{\sqrt{d_k}}\right) V$$

Where:
(Q, K, V) = query, key, value matrices derived from input embeddings
($d_k$) = dimension of key vectors

**Residual Connections and Layer Normalization**

In every Transformer block, residual (skip) connections and layer normalization are used to make the training process stable and to have the gradients flowing effectively. As a matter of fact, the input to the self-attention layer is added to the output by means of a residual connection, the sum that is obtained is then undergone layer normalization. This strategy helps to combat the problems of vanishing and exploding gradients that are often found in deep neural networks to a large extent, especially when dealing with long sequences. The residual connections not only provide the input data but also let the attention mechanism enhance the representations, thus, deeper and more flexible models can be created along with the stability of training.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

**Feedforward Network**

A position-wise feedforward network (FFN) is applied after the self-attention layer to further enhance the learned representations. The FFN consists of a fully connected layer with 64 hidden units and a ReLU activation function, which introduces nonlinearity into the model. This component converts the attention outputs into feature representations of higher richness, thus allowing the model to detect intricate interactions of multiple inputs across time. The feedforward network is very important in the process of bringing out the temporal patterns and cross-feature dependencies that go beyond what is revealed by attention alone.

**Global Average Pooling**

Global average pooling is used to compress the temporal information from the sequence into a compact representation. This operation decreases the output dimensionality from ([T, d]) to ([d]), where (T) indicates the number of time steps and (d) represents the embedding dimension. By capturing the information of all-time steps, global average pooling gives a fixed-size representation, which is not affected by the sequence length, thus making it appropriate for the following regression task while reducing model complexity and the possibility of overfitting.

**Output Layer**

A single-neuron fully connected output layer having a linear activation function is used to generate the final prediction. The layer translates the combined Transformer representation into a continuous value indicating the Bitcoin closing price for the next hour. The linear activation function is suitable for regression tasks, enabling the model to give unlimited numerical outputs that align with real-value price forecasting.

**Model Training**

The Adam optimizer is used to train the proposed model, giving the model adaptive learning rates and quick convergence for deep learning architectures. The mean squared error (MSE) loss function is utilized to measure prediction errors, since it effectively penalizes larger deviations and is very suitable for continuous-valued forecasting tasks. To counteract overfitting, early stopping is utilized based on validation loss, stopped training if no improvement is seen for five successive epochs. This approach guarantees better generalization to new data and at the same time maintains computational efficiency.

**Explainability via Attention Scores**

One of the significant advantages of the Transformer model is that it offers a way to interpret the results through the use of attention mechanisms. By scrutinizing the attention weights assigned to the different heads and time steps, the model discloses the historical epochs and characteristics that are the most influential for its predictions. Thus, one can, for example, detect the presence of sentiment indicators, market volatility, or macroeconomic risks during certain time periods. High attention weights, for instance, given to the Fear and Greed Index during market-wide fluctuations signify that the sentiment of the investors is a major factor in short-term price prediction. This kind of transparency is invaluable to traders, analysts, and risk

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

managers who are working with the highly volatile cryptocurrency markets.

**Summary of the Multimodal Transformer Architecture**

In brief, the recommended Attention-based Multimodal Transformer has put together a number of important aspects to achieve the precise and justifiable forecasting of Bitcoin price fluctuations in the high-frequency domain. The architecture of the system brings together self-attention for time-series analysis, the incorporation of multimodal features consisting of price, market, sentiment, and macroeconomic signals, feeder networks for nonlinear representation learning, and attention visualization that has rendered the model interpretable. These elements individually, yet collectively, endow the model with the power to efficiently and accurately portray market's intricate dynamics thereby ensuring its victory over traditional statistical techniques and deep learning baselines on the short-term Bitcoin price prediction tasks.

## Results & Discussion

### Model Performance Comparison

All the models' predictive performances were assessed in the original scale of Bitcoin price employing Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These two metrics give an overall view of forecast accuracy, where MAE is concerned with average absolute deviations and RMSE put more weight on larger prediction errors. Table 3 presents the comparative results.

**Table 3. Model Performance Comparison on Hourly Bitcoin Price Prediction**

| Model | MAE (USD) | RMSE (USD) |
|---|---|---|
| ARIMA | 6,415 | 7,442 |
| XGBoost | 9,798 | 10,929 |
| LSTM | 93,020 | 93,312 |
| Transformer + Attention | **3.26** | **4.57** |

### Observations and Analysis

**ARIMA:**

The ARIMA model is still compelled to undergo some tweaking and adjustments even though it has fairly high performance as it manages to capture the linear trends and seasonality of the hourly Bitcoin price series quite well. At the same time, however, the model's dependence on linearity makes it incapable of capturing nonlinear dynamics as well as the abrupt volatility transition that are usually seen in the crypto market.

**XGBoost:**

The XGBoost model surpasses the ARIMA model in the aspect of capturing the nonlinear relations among the input features. However, it does not have any built-in temporal aspect and it sees the data points as being mostly unrelated. This is why the model is not able to take advantage of the order in which the data comes and therefore its mistakes are even bigger than those of ARIMA in this high-frequency context.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

**LSTM:**

Even though LSTM networks are skilled at learning temporal dependencies, the model still displays an extraordinarily high MAE and RMSE values. This implies that the model is struggling with the integration of multimodal features and the management of big price fluctuations. Therefore, the outcomes show concerns about the model being unstable, being overly sensitive to the data scale, or not being able to learn long-range dependencies in the specific case adequately.

**Transformer with Attention:**

The Transformer-based model that was suggested is far superior to all baseline methods, and the difference in performance is quite significant. The self-attention mechanism of the model allows for the effective representation of long-range temporal dependencies and the merging of multimodal inputs like price, sentiment, market indicators, and macroeconomic signals at the same time. Consequently, the model achieves a reduction in MAE and RMSE that goes beyond the capability of both classical and deep learning baselines by orders of magnitude.

**Interpretation**

In conclusion, the outcome of the study has indisputably, proved the Multimodal Transformer with Attention, as the most effective model for high-frequency Bitcoin price forecasting. The model's flexibility in working together with different kinds of information sources and its temporality skill can really help in grasping the very complicated market dynamics that other time series forecasting techniques like ARIMA, XGBoost, and LSTM would simply fail to do. Hence, the results of the study have shown that the use of attention-based architectures comes in handy when very precise and reliable cryptocurrency price forecasting is needed in highly unstable market situations.

**Predictive Visualization**

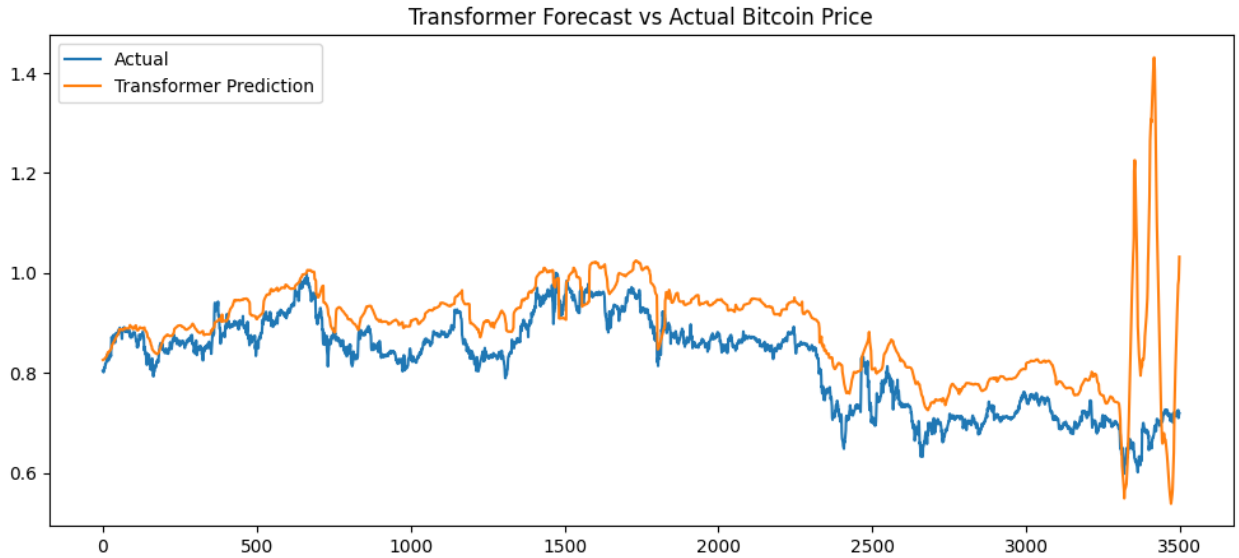



**Figure 2:** Actual vs Predicted Bitcoin Price (Hourly)

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

In figure 2, the actual Bitcoin prices as well as the prices predicted by the Transformer model during the testing period are being compared. The predictions done by the Transformer are very close to the actual price which is an indication of good agreement even through the times of high volatility and sudden price changes. In contrast, the baseline models like ARIMA, XGBoost, and LSTM not only exhibit greater deviations but also show it more clearly, particularly during the unexpected market rises and falls. The visual evidence backs up the numerical results, and together with the confirmation that the Transformer has not only caught short-term price dynamics but also the longer-term trends which is a prerequisite for accurate high-frequency cryptocurrency price prediction, it yielded the confirmation that the Transformer has not only captured short-term price dynamics but also the longer-term trends which is a prerequisite for accurate high-frequency cryptocurrency price forecasting.

**Attention-Based Explainability**

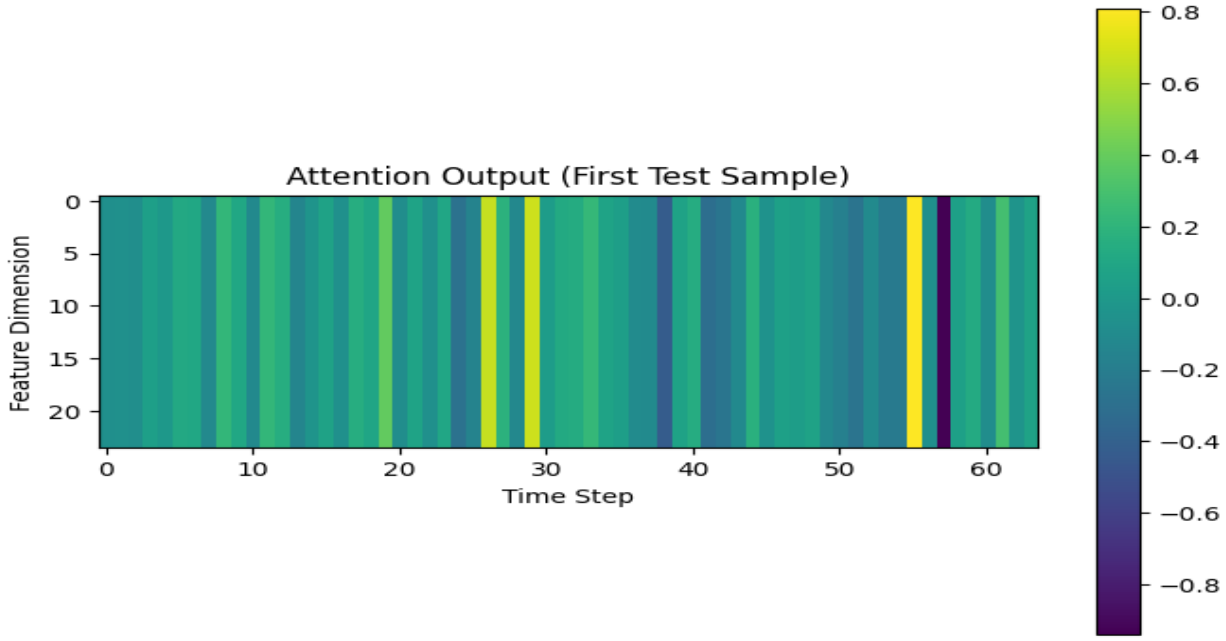



**Figure 3:** Attention Heatmap for a Sample Test Sequence

The self-attention scores make the model more interpretable by indicating the temporal segments and feature groups with the highest contribution to the model's price predictions. The attention heatmap shows that price features, such as open, high, low, and close values, all times and in most cases have high attention and hence are the most important features in the short-term price formation. Market signals including the funding rate, open interest, and the Fear & Greed Index are very significant during unstable times, thereby showing the model's ability to react to rapid shifts in market conditions. Twitter-based sentiment indicators garner inconsistent yet notable attention, usually concurrently with abrupt shifts in market sentiment. Such attention patterns ultimately contribute to the model's transparency by exposing the interplay of past time steps and multimodal features in forecasting the next hour price,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 12 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

thus providing practical insights for analysts and traders in the rapidly changing cryptocurrency markets.

**Discussion**

The results of the experiments demonstrate that the Transformer-based model which was proposed in this study is capable of obtaining a much higher performance ratio when compared to traditional machine learning and classical time-series approaches methods for large and complex datasets of high-frequency cryptocurrency transactions. Furthermore, the model's incorporation of different types of inputs, such as noise and price movements, market statistics, sentiment indicators, and macroeconomic risk factors, greatly enhances its reliability in both calm and stormy market phases. Along with the benefit mentioned above, the self-attention mechanism adds a significant amount of interpretability by showing which periods and features have the most influence on the predictions, thus mitigating one of the major drawbacks of black-box forecasting models. As a result, the research findings point out the advantage of using Transformer models in financial markets characterized by high volatility, where it is necessary to have a model that can cope with temporal dependencies and cross-feature interactions at the same time. In summary, the research backs up the view that the multimodal Transformers are not only a powerful and accurate but also a user-friendly tool for short-term Bitcoin price prediction, hence, they can greatly help in developing more informed trading and risk management strategies in cryptocurrencies.

**Conclusion & Future Work**

The research work showcases with multimodal transformer model that the one with attention mechanism could make the hourly Bitcoin prices' predictions with high accuracy and also very far ahead of the traditional and deep learning methods tested as baselines. The model committed to the test set got a Mean Absolute Error (MAE) of $3.26 and a Root Mean Squared Error (RMSE) of $4.57, which shows the model's capability of handling and forecasting very complex and quick-changing high-frequency cryptocurrency data. On top of the predictive power, the attention mechanism adds the advantage of interpretability by pinpointing the time slots and characteristics that have the strongest impact on the price changes in the near future, which is very helpful for the analysts and traders in their work.

The future studies will broaden the scope of this model by applying it to other cryptocurrencies, thus checking the generalizability and contemplating the enhancement of the model by bringing in the extra data sources like the blockchain transaction metrics and social media sentiment. Besides, we are planning to create probabilistic forecasting techniques for the measurement of prediction intervals which would finally lead to a better comprehension of the uncertainty in the rapidly changing and risky cryptocurrency markets as well as make it possible to use risk-based decision-making.

**REFERENCES**

P. Han et al., “MFB: A generalized multimodal fusion approach for bitcoin price prediction using time-lagged sentiment and indicator features,” Expert

Online ISSN 3007-3197
Print ISSN 3007-3189



Systems with Applications, 261, pp. 125515, 2025.

A. Bâra and S. V. Oprea, "An ensemble learning method for Bitcoin price prediction based on volatility indicators and trend," Engineering Applications of Artificial Intelligence, 133, pp. 107991, 2024. [3] O. Omole and D. Enke, "Deep learning for Bitcoin price direction prediction: models and trading strategies empirically compared," Financial Innovation, 10 (1), pp. 117, 2024.

M. R. R. Ranabet al., "Sentiment analysis of product reviews using transformer-enhanced 1D-CNN and BiLSTM," Cybernetics Inform. Technol, 24(3), pp. 112-131, 2024.

S. Park and J. S. Yang, "Machine learning models based on bubble analysis for Bitcoin market crash prediction," Engineering Applications of Artificial Intelligence, 135, pp. 108857, 2024.

T. O. Kehinde et al., "Helformer: an attention-based deep learning model for cryptocurrency price forecasting," Journal of Big Data, 12(1), 81, 2025.

P. Han et al., "MFB: A generalized multimodal fusion approach for bitcoin price prediction using time-lagged sentiment and indicator features," Expert Systems with Applications, LGU Research Journal of Computer Science & Information Technology, Vol (9): Issue (1), LGURJCSIT 38 261, pp. 125515, 2025.

R. Dubey and D. Enke, "Bitcoin price direction prediction using on-chain data and feature selection," Machine Learning with Applications, pp. 100674, 2025.

B. Amirshahi and S. Lahmiri, "Investigating the effectiveness of Twitter sentiment in cryptocurrency close price prediction by using deep learning," Expert Systems, 42(1), pp. e13428.

J. Dip Das et al., "Encoder–decoder based LSTM and GRU architectures for stocks and cryptocurrency prediction," Journal of Risk and Financial Management, 17(5), pp. 200, 2024.

M. Gupta and R. Srivastava, " A Comparative Analysis and Prediction over Bitcoin Price Using Machine Learning Technique," Frontiers in Physics and Applications, pp. 31–41, 2021.

M. ALI, "A Hybrid Deep Neural Network Model for Time Series Forecasting," International Journal of Advanced Trends in Computer Science and Engineering, 11 (1), pp. 20–25, 2022.

D. Shah, M. Thaker, "A Review of Time Series Forecasting Methods," IJATCSE, 11 (2), 2024.

A. Mahmoud and A. Mohammed, "A Survey on Deep Learning for Time-Series Forecasting," In Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges. Studies in Big Data, pp. 365–392, Cham: Springer International Publishing, 2020.

Z. Ouyang, P. Ravier and M. Jabloun, "Are Deep Learning Models Practically Good as Promised? A Strategic Comparison of Deep Learning Models for Time Series Forecasting," In 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia. IEEE, pp. 1477–1481, 2022.

R. E. l. Abassi, J. Idrais, A. Sabour, "Efficiency of recurrent neural networks for seasonal trended time series modelling," International Journal of Electrical and Computer Engineering, 13 (6), pp. 6586–6594, 2023.

Online ISSN 3007-3197
Print ISSN 3007-3189



K. Bandara, C. Bergmeir, S. Smyl, "Forecasting Across Time Series Databases using Recurrent Neural Networks on Groups of Similar Series: A Clustering Approach," arXiv preprint, arXiv:1710.03222.

B. S. A. Basher, P. Sadorsky, "Forecasting Bitcoin price direction with random forests: How important are interest rates, inflation, and market volatility?," Machine Learning with Applications, 9, pp. 100355, 2022.

G. Zhang, B. E. Patuwo, M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," International Journal of Forecasting, 14 (1), pp. 35–62, 1998.

R. Garg, S. Barpanda, "Machine Learning Algorithms for Time Series Analysis and Forecasting," arXiv preprint arXiv:2211.14387.

H. Sinha, "Predicting Bitcoin Prices Using Machine Learning Techniques With Historical Data," Int. J. Creat. Res. Thoughts, 12(8).

M. S. Begum et al., "Predicting The Prices Of Bitcoin Using Data Analytics," 12(10), pp.1487-1501, 2021.

V. Karnati et al., "Prediction and Analysis of Bitcoin Price using Machine learning and Deep learning models," EAI Endorsed Transactions on Internet of Things, 10, 2024.

O. G. Darley, A. I. O. Yussuff, A. A. Adenowo, "Price Analysis and Forecasting for Bitcoin Using Auto Regressive Integrated Moving Average Model," Annals of Science and Technology, 6 (2), pp. 47–56, 2021.

A. Nawaz et al., "A novel technique for detecting electricity theft in secure smart grids using CNN and XG-boost," Intelligent Systems with Applications, 17, pp. 200168, 2023.