https://doi.org/10.1186/s40537-025-01135-4

Journal of Big Data

RESEARCH Open Access

Helformer: an attention-based deep learning model for cryptocurrency price forecasting

T. O. Kehinde[1*], Oluyinka J. Adedokun[2], Akpan Joseph[3], Kareem Morenikeji Kabirat[4], Hammed Adebayo Akano[5] and Oludolapo A. Olanrewaju[3]

*Correspondence: temitope.kehinde@connect.polyu.hk

[1] Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
[2] Department of Industrial & Systems Engineering and Engineering Management, University of Alabama in Huntsville, Huntsville, USA
[3] Department of Industrial Engineering, Durban University of Technology, Durban, South Africa
[4] Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria
[5] School of Life and Environmental Science, Deakin University, Geelong, Australia

Abstract

Cryptocurrencies have become a significant asset class, attracting considerable attention from investors and researchers due to their potential for high returns despite inherent price volatility. Traditional forecasting methods often fail to accurately predict price movements as they do not account for the non-linear and non-stationary nature of cryptocurrency data. In response to these challenges, this study introduces the Helformer model, a novel deep learning approach that integrates Holt-Winters exponential smoothing with Transformer-based deep learning architecture. This integration allows for a robust decomposition of time series data into level, trend, and seasonality components, enhancing the model's ability to capture complex patterns in cryptocurrency markets. To optimize the model's performance, Bayesian hyperparameter tuning via Optuna, including a pruner callback, was utilized to efficiently find optimal model parameters while reducing training time by early termination of suboptimal training runs. Empirical results from testing the Helformer model against other advanced deep learning models across various cryptocurrencies demonstrate its superior predictive accuracy and robustness. The model not only achieves lower prediction errors but also shows remarkable generalization capabilities across different types of cryptocurrencies. Additionally, the practical applicability of the Helformer model is validated through a trading strategy that significantly outperforms traditional strategies, confirming its potential to provide actionable insights for traders and financial analysts. The findings of this study are particularly beneficial for investors, policymakers, and researchers, offering a reliable tool for navigating the complexities of cryptocurrency markets and making informed decisions.

Keywords: Helformer, Cryptocurrency forecasting, Bitcoin, Transformer, Neural networks, Time series

Introduction

The cryptocurrency domain has received growing attention from investors, regulators, fund managers, policymakers, and researchers since its first coin, Bitcoin (BTC), which was initially launched in 2008 by an anonymous individual or group of individuals called Nakamoto [40]. Its growing popularity, which increased from zero worth at the time of launch in 2009 to the all-time highest price of 103,900.47 USD on 5th December 2024, is due to its appealing features such as Proof-of-Work and Proof-of-Stake, consensus algorithm, and secured ledgers [53], which are different from conventional financial assets such as gold, bonds, physical currency, and stocks. Its worth is based on the confidence of its underlying innovative algorithms, such as traceability and decentralization [34, 36], rather than any tangible asset, making it independent of regulation, manipulation, government interference, and policy changes. It also has intrinsic characteristics such as low transaction costs and secure peer-to-peer (P2P) payment [3].

Many studies have recognized cryptocurrencies as an investment asset. In this regard, some recent research has explored the potential synergies between cryptocurrencies and other investment assets such as gold, commodities, stocks [29, 30],Kehinde, Chung, et al., 2023), and physical currencies. Some existing research provides empirical evidence demonstrating that cryptocurrencies exhibit a low correlation with traditional financial assets. Consequently, this characteristic positions cryptocurrencies as a valuable hedge in investment portfolios [27]. Consequently, BTC, which is the first, most valued, and most popular coin, has been emphasized to allow hedging investment strategy against other investment assets such as gold, oil, stocks, and commodities due to high return and low correlation with other investment assets [58]. As of March 2025, there are more than 10,700 active and valuable cryptocurrencies, with over 420 million users worldwide. Out of the active cryptocurrencies available, only the top 20 accounts for nearly 90% of the total market, with around 251 spot exchanges and a total market cap of 2.54 trillion USD (https://coinmarketcap.com, accessed on 11th March 2025).

Due to the huge returns associated with trading cryptocurrency, it is worth noting that it comes with high risk because of the large price fluctuations commonly experienced in trading, as it is always traded online in real-time, traded round the clock with no official opening or closing time. In this connection, most people involved in this kind of trading are usually experienced traders and algorithm trading bots. It is estimated that more than half of the trading volume is typically traded by bots, and these bots require robust deep-learning models to analyze, predict, and make successful trades [7]. Given the volatile nature of cryptocurrencies, it is crucial for investors to accurately predict cryptocurrency prices to manage risks, diversify their portfolios, and maximize returns. Effective prediction strategies and algorithms can significantly guide investors in making both short and long-term investment decisions.

In the past, different cryptocurrency price forecasting methods have been developed, and these can be categorized into statistical, machine learning, and deep learning methods. Early work in this area focuses on traditional statistical techniques, whereas ARIMA is the most commonly used conventional method among these techniques [1]. However, these approaches only assume time series to be linear, which is usually not applicable to assets like cryptocurrency, especially when dealing with an extensive dataset that spans various periods such as the pandemic period (e.g., COVID-19 pandemic), war period (e.g., Russia-Ukraine war and Israel-Hamas war). Furthermore, another limitation of the statistical model is the assumption of normal distribution of variables, which is unrealistic for chaotic and non-stationary data like cryptocurrency. To this effect, machine learning methodology was introduced to overcome these limitations.

Machine learning methods are designed to extract the non-linear nature inherent in large datasets of the cryptocurrency market. Although early machine learning models like Linear regression and Logistic regression seem to be parametric, later models such as Support Vector Machine (SVM), k-Nearest Neighbors (KNN) [51], and Multi-Level Perceptron (MLP) are non-parametric and do not require a prior understanding of the distribution of data to model the non-linear relationship among variables. However, one of the limitations of using machine learning is that they are susceptible to overfitting, especially when handling long sequence time series forecasting (LSTF) data such as cryptocurrency data. Another limitation is that their models produce a more considerable error, making the model perform poorly when subjected to trading strategy. In this regard, deep learning was later introduced to explore and overcome the weakness of machine learning models.

With its capacity to outperform statistical and machine learning models, deep learning is created to explore intricate patterns of more complex data. These models have shown exceptional performance in handling complex data, and subsequently, models such as Recurrent Neural Network (RNN) and its variants are designed to model LSTF where the order of data is a priority. RNN has shown good performance in modelling time series data; however, the problem of vanishing gradient or exploding gradient has been the limitation of this model when handling LSTF data, which, in turn, leads to the development of more variants of its kind, including Long Short-Term Memory (LSTM), Bi-directional LSTM (BiLSTM), and Gated Recurrent Unit (GRU). Though LSTM has been proven to be the most used time series model, some researchers have shown that BiLSTM and GRU can surpass the accuracy of LSTM in some instances and for some data. Despite the success recorded by RNN and its variants in making accurate predictions, its computation still suffers complexity due to the sequential processing that is inherent in these models. In this connection, more research has been done on building models that can compute in parallel and generate exceptional outputs.

Vaswani et al. [59] proposed a Transformer neural network, an NLP-focused paradigm, to address serial computation and model complexity. The authors presented the Transformer model, which uses self-attention. This invention differs from existing approaches that mainly used recurrence or convolutions. Transformer computes various attention scores simultaneously, allowing it to focus on different sequence parts and improve context understanding. Unlike other models, Transformer captures linkages and dependencies inside word vectors regardless of distance. Instead of sequential processing, the Transformer allows for better parallelization during training, improving speed relative to all other models, especially for massive data. Transformer neural networks have achieved breakthroughs in image processing, speech processing, language translation, computer vision, healthcare and bioinformatics, robotics, and autonomous designs. However, their use in LSTF, such as the cryptocurrency market, is still early. Also, it is worth noting that many researchers have proved that cryptocurrency data possess attributes like non-stationary and seasonality,meanwhile, traditional neural networks like Artificial Neural Networks (ANN), RNN, LSTM, GRU, and Transformer are not designed to handle these complexities, leading to poor predictions.

Inspired by the work of Smyl [55], which decomposes time series into trend and seasonal parts, this work introduces a new variant of Transformer called Helformer, which has been designed to handle complex data that exhibit non-stationarity and seasonality. The suggested model uses Holt-Winters exponential smoothing to extract level, trend, and seasonality from a series decomposition method. This breakdown strategy helps the attention mechanism grasp global trends efficiently. The conventional Transformer model uses positional encoding coupled with input embedding to turn high-dimensional word vectors into low-dimensional ones for NLP applications. This study, a non-NLP problem, uses an LSTM component to substitute a Feed Forward Network (FFN) mechanism in the Encoded architecture to capture temporal dependencies, an attribute inherent in time series forecasting. This work uses only the encoder component, as Haryono et al. [24] supported the claim that using a single encoder component is more effective than using dual components, especially in time series prediction, because it reduces memory complexity and computational demand.

Although there is a continuous rise in the weekly debut of new coins, developing separate models for individual models may be time-consuming and resource-demanding. As observed in previous works, most studies, investors and traders focus on four notable coins: BTC, Litecoin (LTC), Ethereum (ETH), and Ripple (XRP) [8, 19, 43, 60, 66]. With over 10,700 active cryptocurrencies and the possibility of new debuts periodically, developing a model for each cryptocurrency is quite challenging. The transfer learning technique capitalizes on the accumulated insights from pre-trained model iterations, using them as a foundation for tackling novel tasks. This transfer learning technique allowed the model to effectively generalize across different cryptocurrencies, showcasing its potential for broader applications in cryptocurrency markets. Unlike previous studies, this work intends to build its novel model on BTC data and test its generalization and cross-learning ability on other selected cryptocurrencies. In addition, since a good model may not demonstrate a viable trading strategy, unlike previous studies, this work designs a simple trading strategy to evaluate the feasibility of the proposed model to make a profitable investment. It is worth noting that the proposed Helformer model is developed alongside other sophisticated deep learning models to serve as benchmarks. The robustness of the Helformer model is tested by doing a comparative analysis with notable existing studies to demonstrate the reliability of Helformer in outperforming existing works. The contributions of this work are as follows:

1. A novel model is designed to predict highly volatile assets like cryptocurrency.
2. Unlike previous studies that frequently use manual tuning for machine learning models, this work implements Bayesian optimization with Optuna for hyperparameter tuning to generate robust predictions.
3. Empirical analysis shows minimal errors and exceptional performance, outperforming all existing state-of-the-art methods and studies.
4. This work is the first implementation of the Helformer model, the validation of which was tested across 15 cryptocurrencies.
5. Last, this work showcases the practical implications and potential profitability of targeted cryptocurrencies to generate substantial returns.

The remaining sections of this work are systematically structured as follows: Sect. "Related research" gives a summary of existing studies on cryptocurrency prediction. Sect. "Methodology" describes the methods and framework adopted in this study. Sect. "Empirical results and discussions" discusses empirical results, while

Sect. "Conclusion, Limitations, and Future directions" serves as the final part of the work, summarizing the acquired insights and outlining a direction for future works.

Related research

This section reviews past and current advances in cryptocurrency price forecasting. Further, it categorizes existing studies into three types: classical, machine learning, and deep learning approaches.

Cryptocurrency

The use of cryptocurrency for financial transactions has increased in the last decade. In this regard, several countries, including Ukraine, El Salvador, Japan, South Korea, the United States, Switzerland, Germany, Portugal, Malta, and UAE, have legalized its usage as a legal payment method [35, 66]. Empirical evidence suggests that the predictability issues of cryptocurrency are related to attributes such as: heavily tailed distributions of cryptocurrency returns, autocorrelations for relative and absolute returns exhibiting different decay rates, strong leverage effect and volatility clustering, and power-law correlation between price and volatility. These features contribute to the predictability issues of cryptocurrency. Ideally, most assets are generally predicted by technical analysis, financial analysis, or a combination of both. However, due to the decentralized nature of cryptocurrency, its electronic generation, and its newest to the financial world, predicting cryptocurrencies has been challenging because they are unrelated to any fundamentals, and market sentiments mainly influence them [33, 48]. In this realization, past works have explored approaches such as classical, machine learning, and deep learning in predicting cryptocurrency prices, returns, and volatilities.

Classical approach to cryptocurrency price forecasting

This approach comprises statistical models, such as Moving Average, AutoRegressive Moving Average (ARMA), AutoRegressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), Generalized AutoRegressive Conditional Heteroskedasticity (GARCH), Exponential Smoothing, which have long been applicable in time series [14]. These models are based on statistical theory and are efficient in certain market scenarios, especially when the market exhibits linear predicted patterns or trends. Generally, classical models marked a notable progression in cryptocurrency prediction, especially when predicting cryptocurrency volatility. For instance, Conrad et al. [12] explore the volatility components of cryptocurrencies, particularly BTC, using the GARCH-MIDAS model. The study investigates the influence of macroeconomic and financial factors on both short-term and long-term BTC volatility. The results suggest that BTC's volatility is unique compared to other financial assets, behaving pro-cyclically and responding differently to economic conditions. Similarly, Walther et al. [60] examine the impact of various macroeconomic and financial factors on the volatility of major cryptocurrencies, including BTC, LTC, ETH, XRP, XLM, and the CRIX cryptocurrency index. Using the GARCH-MIDAS framework, the authors differentiate between short-term and long-term volatility components and identify the most influential exogenous drivers.

Catania et al. [8] investigated the predictability of cryptocurrency time series, particularly focusing on BTC, LTC, XRP, and ETH. The authors compare a variety of univariate and multivariate VAR models for point and density forecasting, utilizing dynamic model averaging (DMA) and dynamic model selection (DMS) to combine and select among these models. Notably, the popularity of all the aforementioned classical models stems from their simplicity and interpretability,however, they frequently fail to capture the non-linear nature, non-stationary nature, and intricate complexities associated with the cryptocurrency market. This limitation occurs due to their dependence on linear assumptions regarding market behaviour. This gap has resulted in an increasing trend towards using more advanced techniques like machine learning that can effectively handle the non-linear and non-stationary nature of the cryptocurrency market.

Machine learning approach to cryptocurrency price forecasting

Driven by cryptocurrency's highly volatile and non-linear nature, attention has been shifted to applying machine learning, which can analyze large volumes of data, identify patterns, and adapt to dynamic market conditions. Machine learning models can reveal complex patterns in data that may not be immediately obvious, providing a more sophisticated comprehension of market dynamics compared to conventional statistical models. In this realization, some researchers have already employed machine learning approaches such as Logistic Regression, KNN, Decision Tree, SVM, and many more to develop prediction models capable of generating super profits. In addition, to generate more robust predictions, while some researchers have employed ensemble models, including Random Forest, AdaBoost, XGBoost, CatBoost, and LightGBM, others have engaged in hybrid models to predict cryptocurrency prices, returns, and volatilities.

Existing studies already confirmed the robustness of machine learning models such as ANN to outperform classical models. For instance, Nakano et al. [41] investigated the application of ANNs for predicting BTC returns based on high-frequency trading data. The authors utilize a seven-layer ANN model that processes technical indicators calculated from BTC historical data at 15-min intervals to identify potential trading signals. Their approach significantly outperforms a traditional buy-and-hold (B&H) strategy, particularly during periods of high volatility, such as from December 2017 to January 2018, when BTC experienced substantial losses. In another study, Kristjanpoller and Minutolo [32] propose a hybrid framework combining GARCH models, ANN, technical analysis indicators, and Principal Component Analysis (PCA) for forecasting the volatility of BTC. The authors argue that while traditional GARCH models capture certain aspects of financial time series volatility, integrating them with ANN and technical indicators such as the Relative Strength Index (RSI) and Moving Average Convergence Divergence (MACD) enhances predictive performance. Ibrahim et al. [26] compare various time-series modelling methods for predicting BTC price movements in short timeframes. The study finds that the MLP achieves the highest accuracy at 54%, outperforming several models but only slightly better than a simple momentum strategy.

Moving forward, Rathore et al. [50] explore the challenges of predicting BTC prices due to their volatility and dynamic trends. The authors compare traditional models like ARIMA and LSTM, noting their limitations in handling seasonality and outliers, and propose the use of the Facebook Prophet model for better handling of time series data. The model is designed to account for seasonality and outliers, making it more suitable for real-world cryptocurrency predictions. The study demonstrates that the Prophet model yields more accurate results compared to Naïve and other traditional models. For robust predictions, many researchers have explored the possibility of using ensemble models for cryptocurrency forecasting. For instance, Sun et al. [56] apply a Light Gradient Boosting Machine (LightGBM), a machine learning algorithm. The study finds that the LightGBM model outperforms traditional models such as SVM and RF in terms of robustness and forecasting accuracy, particularly in medium-term predictions (e.g., 2-week periods). Next, using machine learning techniques, Sebastião and Godinho [54] investigate the predictability and profitability of trading strategies for three major cryptocurrencies: BTC, ETH, and LTC. The study spans from August 2015 to March 2019, a period marked by significant market fluctuations, including bull and bear markets. The authors employ multiple machine learning models, including linear models, RF, and SVM, to forecast cryptocurrency returns based on trading and network activity data. The findings reveal that although individual models' performance can vary under changing market conditions, ensemble models, particularly ones requiring consensus, show robust profitability.

More recently, the work of Chang et al. [10] put forth a model for forecasting cryptocurrency price using a combination of Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), time series clustering, and reconstruction of intrinsic mode functions (IMFs). The scheme decomposes the BTC price into IMFs using CEEMDAN, then groups these IMFs into three clusters using a robust ensemble clustering approach. The results of this approach demonstrate significant improvements compared to traditional and more straightforward models. Although machine learning methods are proficient in modelling non-linear connections and extracting insights from complex datasets, they are susceptible to overfitting, especially when handling LSTF data such as cryptocurrency. Consequently, investors and researchers increasingly switch to state-of-the-art approaches, such as deep learning models.

Deep learning approach to cryptocurrency price forecasting

Deep learning models are expected to provide a more thorough predictive ability in the highly volatile cryptocurrency market. The exponential increase in computational capacity in recent years has accelerated the emergence of deep learning methodologies, fundamentally transforming diverse financial domains, such as the cryptocurrency market. Deep learning, a kind of machine learning distinguished by its utilization of multilayered neural networks, has significantly transformed various domains, such as finance. The emergence of deep learning models, such as Convolutional Neural Networks (CNNs), RNNs, LSTMSs, and GRU, signifies the most recent frontier in forecasting cryptocurrency prices, returns, and volatilities, as they exhibit their outstanding performance in capturing temporal dependencies and non-linear correlations.

RNN has shown good performance in modelling time series data; however, the problem of vanishing gradient or exploding gradient has been the limitation of this model when handling long time series data, which, in turn, leads to the development of more variants of its kind which include LSTM, BiLSTM, and GRU. Though LSTM has been proven to be the most used time series model, some researchers have shown that BiLSTM and GRU can surpass the accuracy of LSTM in some instances and for some currencies. For example, Hamayel and Owda [21] developed three models, LSTM, GRU, and Bi-LSTM, to predict the prices of cryptocurrencies such as BTC, ETH, and LTC. The study finds that the GRU model provides the most accurate predictions with the lowest error. Similar results were achieved in a similar experiment performed by Dutta et al. [15], Hansun et al. [23], and Jin and Li [28]. In contrast, Seabe et al. [53] repeated a similar experiment with a contrary result where Bi-LSTM outperforms the GRU model. More recently, Golnari et al. [19] presented a novel deep learning approach for predicting cryptocurrency prices, focusing specifically on BTC. The authors propose a Probabilistic GRU (P-GRU) model that incorporates probabilistic features to provide a probability distribution for predicted values, improving prediction accuracy under volatile market conditions. The model's performance was compared with other established models, including GRU, LSTM, and their probabilistic variants, using 1 year of BTC price data sampled at 5-min intervals. The P-GRU model outperformed the traditional models in accuracy and robustness.

Empirical evidence from numerous studies indicates that hybrid models consistently outperform singular models, suggesting that they offer superior performance moving forward. As an example, Zhong et al. [63] introduce a hybrid model LSTM-ReGAT for predicting cryptocurrency price trends by leveraging individual cryptocurrency features and their interrelations. The hybrid model combines LSTM networks for capturing time series patterns and a Relation-wise Graph Attention Network (ReGAT) to utilize the interrelationships between cryptocurrencies. The model builds a cryptocurrency network using shared features like technology, industry, and investor co-attention. This network-centric approach is validated using real-world data, showing that LSTM-ReGAT outperforms traditional models in both prediction accuracy and profitability in trading simulations for BTC and cryptocurrency portfolios. Other notable studies whose work demonstrates the exceptional performance of hybrid deep learning models against straightforward models include Patel et al. [46], Nasirtafreshi [42], Goodell et al. [20], and Girsang [18]

CNN, which has been traditionally used in image processing, has shown exceptional performance when used as a feature extraction mechanism in hybrid models for cryptocurrency prediction. For example, Alonso-Monsalve et al. [3] explore the effectiveness of CNN and hybrid CNN-LSTM models in predicting high-frequency cryptocurrency price trends. The authors compare four neural network architectures: CNN, hybrid CNN-LSTM, MLP, and Radial Basis Function Neural Network (RBFNN), to classify whether six common cryptocurrencies will increase in value against USD in the next minute. Using eighteen technical indicators derived from 1 min resolution exchange rate data over one year, the study shows that the CNN-LSTM models outperform the others significantly, thus emphasizing their advantages over traditional machine learning methods in high-frequency trading scenarios. In a similar vein, Cavalli and Amoretti [9] present a novel approach for predicting BTC price trends using a One-Dimensional CNN (1D CNN) model. The authors propose a comprehensive methodology that integrates BTC historical values, financial indicators, social media sentiment analysis from Twitter, and blockchain transaction data to create extensive datasets for model training. The study introduces a cloud-based system with an efficient distributed architecture to handle large data collection and preprocessing tasks. Experimental results show that the proposed 1D CNN model outperforms traditional LSTM models in predicting BTC trends, achieving higher accuracy rates. Other notable studies that demonstrate CNN incorporation in their hybrid models include Livieris et al. [37], Zhang et al. [62], and Peng et al. [47]

Some recent works perform comparative studies of various models, including classical, machine learning, deep learning, ensemble, and hybrid models, to determine which is exceptional. Notable works in these categories include Oyedele et al. [44] and Bouteska et al. [7]. However, most deep learning models are not equipped with attention mechanisms to process tasks in parallel, making them prone to complexity in learning more challenging temporal patterns. In this regard, consideration has been shifted to exploring attention-based related models in modelling LSTF tasks in order to explore this domain of knowledge..

Attention-based approach to cryptocurrency price forecasting

To overcome the limitation of serial computation and model complexity as frequently experienced in existing deep learning models, Vaswani et al. [59] put forth a model called Transformer. The authors proposed the Transformer model, which relies entirely on self-attention mechanisms. The fundamentals of the Transformer are the self-attention mechanism and multi-head attention, and these enable the model to assess the importance of different words in a sequence through the use of multi-head attention while processing each word. In addition, it computes multiple attention scores in parallel, giving room to concentrate on diverse parts of a sequence concurrently and enhancing its ability to understand the context. Unlike traditional models, this enables capturing relationships and dependencies regardless of distance within word vectors. Since the Transformer does not rely on sequential computation, it allows for greater parallelization during training, leading to significant speed improvements compared to all existing models, especially when dealing with big data. Although Transformer neural networks have successfully made unprecedented results in many domains such as image processing, speech processing, language translation, computer vision, healthcare and bioinformatics, robotics, and autonomous designs, their application in LSTF, such as the cryptocurrency market, is in its early stage. Figure 1 depicts a typical architecture of the Transformer model.

Recent applications of Transformer neural networks to cryptocurrency include the works of Tanwar and Kumar [57] and Amadeo et al. [4]. Tanwar and Kumar [57] explore a hybrid approach to predict cryptocurrency prices by integrating Transformer models and LSTM networks. The study focuses on forecasting the prices of major cryptocurrencies like BTC, ETH, and Binance Coin (BNB). The authors first apply Multifractal Detrended Fluctuation Analysis (MFDFA) to process the time-series data, capturing both short and long term temporal dependencies. The hybrid model leverages LSTM's ability to retain temporal information and the Transformers' self-attention mechanism for better prediction accuracy. Further, Amadeo et al. [4] explore the use of the Temporal Fusion Transformer (TFT) model for predicting BTC prices across multiple future time steps. The authors highlight the significant price volatility of BTC and the challenges associated with accurate forecasting. Since the Transformer model was introduced to be successful in other domains, its application to the time series model is limited by three points, as suggested by Zhou et al. [64]

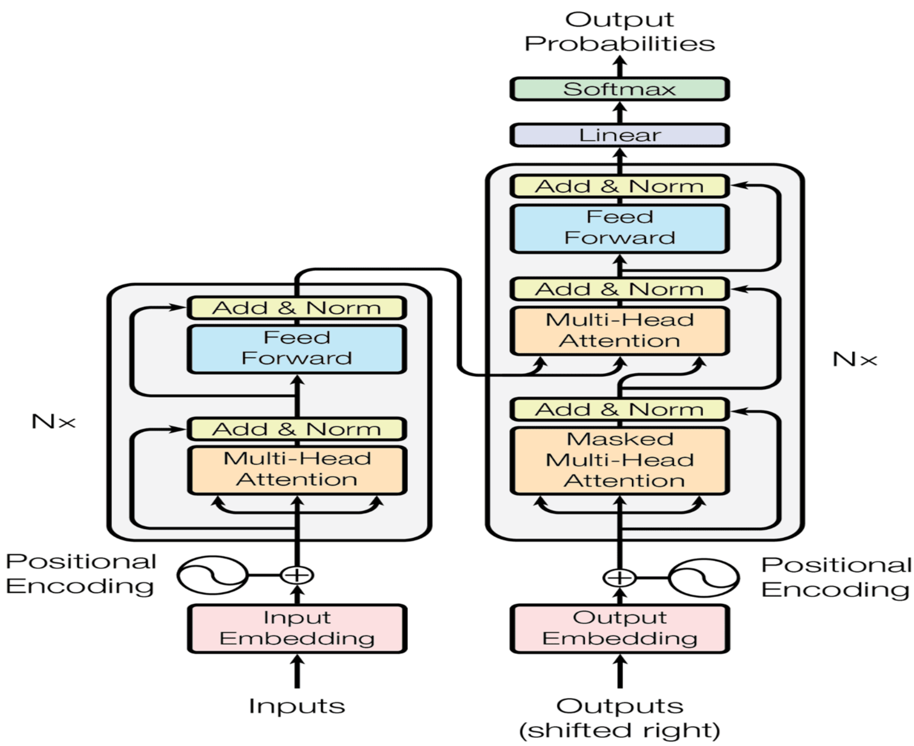

Fig. 1 Transformer model configuration [59]

and Lu et al. [38]. These limitations include significant time complexity and memory consumption, scalability challenges, and decreased processing performance for lengthy outputs. These problems can impede its direct implementation in LSTF for structured datasets.

Several variants of the Transformer model have been developed to address these inherent limitations. These variants include Autoformer, Informer, FDG-Trans, FEDFormer, Sparse Transformer, LogSparse Transformer, Longformer, Reformer, Performer, RSMformer, and many more [5, 28, 61, 65]. Conversely, none have been applied to significantly improve predictions, especially in a highly volatile cryptocurrency market. Also, it is worth noting that many researchers have proved that cryptocurrency data possess attributes like non-stationary and seasonality; meanwhile, traditional neural networks are not designed to handle these complexities, leading to poor predictions. Inspired by the work of Da Silva et al. [13], Li et al. [35], Fallah et al. [16], Ghosh et al. [17], and Koo and Kim [31], which decomposes time series before applying neural networks, this work establishes a new variant of Transformer called Helformer, which has been designed to handle complex data that exhibit non-stationarity and seasonality. Helformer differentiates itself from earlier models by automatically learning and extracting seasonal patterns directly from the time-series data instead of relying on manually developed dynamic time-dependent variables. This feature enables enhanced and simplified pattern identification without requiring manual input on time-dependent variables. The proposed model is trained using Bayesian optimization and tested for transfer learning ability, allowing it to predict the performance of other cryptocurrencies by leveraging the knowledge gained from saved weights of the previously learned model.

Methodology

This section discusses the proposed model, data collection, data preprocessing, model development, systematic framework, experimental settings, and all other requirements for a successful model implementation.

Helformer

Previous studies, including Da Silva et al. [13], Li et al. [35], Jin and Li [28], Fallah et al. [16], Ghosh et al. [17], and Koo and Kim [31], have extensively examined the trend and seasonality in cryptocurrency markets. These researchers employed decomposition methods such as Singular Spectrum Analysis (SSA), Empirical Mode Decomposition (EMD), and Variational Mode Decomposition (VMD) to analyze the data. This decomposition is crucial for enhancing neural networks, which typically lack inherent parameters to account for the levels and seasonality of time series data, as noted by Koo and Kim [31]). However, despite the use of decomposed-based neural networks in these studies, significant prediction errors persist. This highlights the ongoing need for research aimed at developing more robust and sophisticated models to address these challenges.

The proposed Helformer uses a single encoder structure instead of the dual components proposed in traditional Transformer architecture. This encoder structure of Helformer consists of a series decomposition block, an attention mechanism, residual connections, an LSTM component, and a dense layer. Using just a single structure of Transformer architecture reduces the model complexity memory bottlenecks and reduces computational resource usage without compromising prediction accuracies [24]. The Helformer model is designed to predict the closing price of BTC for the next trading day based on a specified window size. The proposed model incorporates the Holt-Winters exponential Smoothing method with a modified transformer-based architecture optimized using Optuna. Initially, the Holt-Winters smoothing layer is employed to decompose the BTC closing price data into its level, trend, and seasonal components. This decomposition allows for a better understanding and removal of seasonality from the data, resulting in a deseasonalized dataset that improves the model's predictive capability. The normalized data is then used as input for the multiple attention blocks and an LSTM layer. The attention blocks in the model enable it to focus on significant features within the data, while the LSTM layer captures the temporal dependencies essential for accurate time-series forecasting. The model is further optimized using Optuna, which fine-tunes hyperparameters such as learning rate, dropout rate, and the number of attention heads, ensuring the best possible performance. Additionally, the exponential smoothing coefficients are directly incorporated into the neural network model, which enables them to be improved with other parameters within the same model optimizer.

The decomposition block uses Holt-Winters smoothing to pinpoint crucial parameters. These are known as local parameters: alpha (α) and gamma (γ) whose value ranges between 0 and 1. As detailed in Eq. 1 and Eq. 2, it decomposes the inputs into seasonality (S_t) and level (L_t) components at every data point (X_t) before being fed into the multi-head attention mechanism whose role is to study the complex, non-linear and non-stationary pattern of the smoothed data to extract the trend component and dependencies. Equation 1 computes a weighted mean by blending the seasonality with the level-adjusted observations from the previous time point (t-1), while Eq. 2 forecasts the seasonal component as a weighted mean for a future time point (t + m). It predicts the seasonality component (X_t/L_t) based on the past estimate (S_t); meanwhile, the deseasonalization is conducted using Eq. 3.

$$L_t = \alpha \frac{X_t}{S_t} + (1 - \alpha)L_{t-1} \tag{1}$$

$$S_{t+m} = \gamma \frac{Y_t}{L_{t+1}} + (1 - \gamma)S_t \tag{2}$$

$$Y_t = \frac{X_t}{S_t L_t} \tag{3}$$

The integration of the multi-head attention mechanism with the decomposition block in the proposed model transcends the mere ensemble combination of exponential smoothing and neural networks; it synchronizes the fitting of all parameters with the neural network weights concurrently. This model processes sequential data that has been refined to eliminate irrelevant information and seasonal variations, rendering it more suitable for the attention mechanism. As illustrated in Fig. 2, the multi-head attention mechanism engages with the smoothed data by analyzing all its components in parallel rather than in a sequential manner. This parallel data processing ability allows the model to recognize global dependencies across the entire input series effectively. Such a strategy significantly enhances the speed of the training process compared to traditional methods, which process data points one at a time. Typically, the self-attention configuration of the Transformer model is outlined in Eq. 4.

$$Attention(Q, K, V) = soft\max(\frac{QK^T}{\sqrt{d}})V \tag{4}$$

where d is the hidden dimension of the keys. The matrices $Q, K, V \in \mathbb{R}^{T*d}$ represent the query, key, and value matrices, respectively. These matrices are the outputs of three distinct linear layers that share the same input. The self-attention mechanism offers a novel approach to concentrate on crucial local information.

Nonetheless, employing multiple self-attention mechanisms, known as multi-head attention, can enhance performance. Within this framework, each attention function operates simultaneously, processing the corresponding projected versions of the query, key, and value matrices. The outputs of all these attention functions are then amalgamated through concatenation and subsequently transformed into the final output via a linear layer. The formula for multi-head attention is encapsulated in Eq. 5.

$$\begin{aligned} &MultiHead(Q, K, V) = Concat(head_1, head_2, ..., head_h)W^O \\ &head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{5}$$

where, $i = 1, ..., h$ and $W_i^Q,\ W_i^K,\ W_i^V$ are weights of networks.

Going forward, the add & norm layers are added as they are critical in stabilizing the training process and improving model performance. The incorporation of the add & norm layer in the Helformer model greatly improves stability and speed in

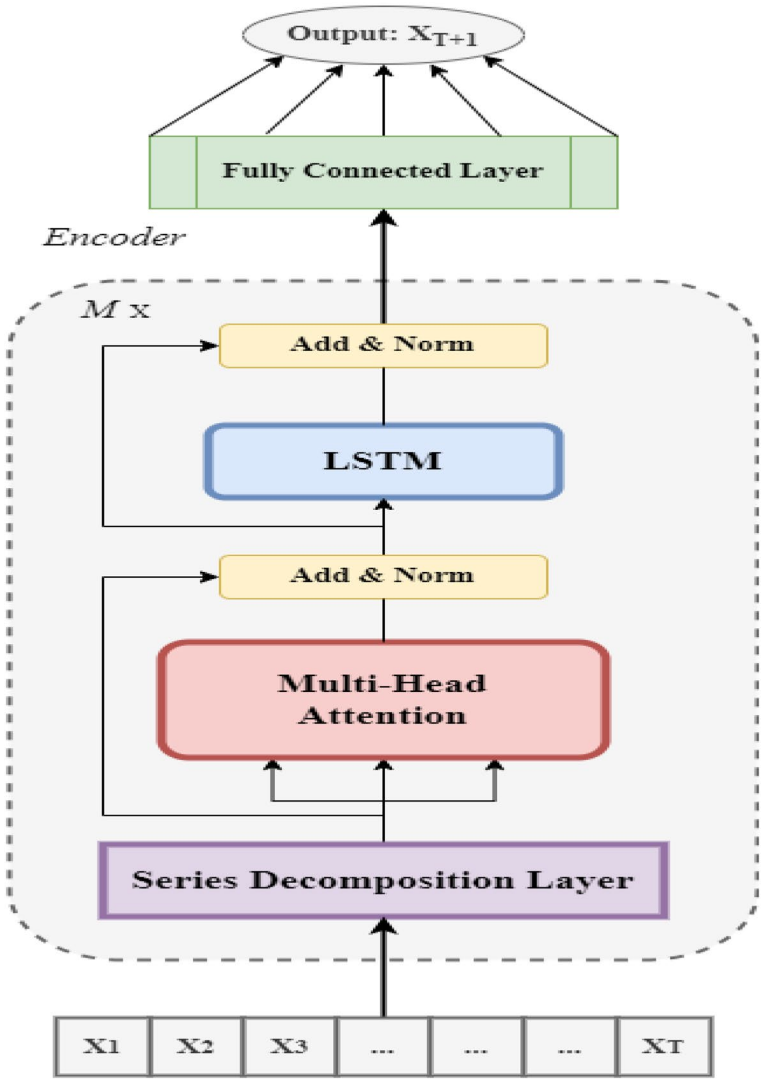

Fig. 2 Helformer architecture

the training process. The addition component utilizes residual connections, effectively addressing the issue of vanishing gradients by enabling the direct transfer of gradients through the layers. Subsequently, the normalizing procedure employs layer normalization to equalize the output across features. This is essential for ensuring a uniform scale that promotes accelerated and stable training. This combination not only simplifies the learning process but also guarantees that the model adjusts rapidly and efficiently to the intricacies of the input data. Also, an LSTM layer was introduced to replace the conventional FFN typically employed in regular transformers. The LSTM layer captures the temporal dependencies essential for accurate time-series forecasting. This design, as depicted in Fig. 2, presents the proposed architecture of the Helformer model.

Data

Data collection

In this work, the proposed model is trained using the dataset of the most popular and most valued cryptocurrency, BTC. As cryptocurrencies are traded round the clock with no specific opening or closing times, the closing price data used in this analysis are taken at midnight (12:00 am) each day, marking the end of the trading day. Afterward, the model leverages the pre-trained BTC model to forecast prices for 15 other active top cryptocurrencies in the decreasing order of their market cap while excluding stablecoins. This technique allowed the model to effectively generalize and perform cross-learning across different cryptocurrencies, showcasing its potential for transfer learning. The daily closing prices for all the selected cryptocurrencies analyzed in this study were downloaded from Yahoo Finance on 21st July 2024. The number of samples varies for each currency, as these coins have different launch dates; therefore, datasets were downloaded based on the maximum period available in the chosen database. Yahoo Finance was selected as a data source due to its reputation and reliability in maintaining accurate and dependable data over time, as well as its widespread use in numerous notable studies. Table 1 presents the details of the collected data along with their basic statistical analysis. It provides an overview of the collected data, including the number of samples, the start and end dates for the data collection period, and basic statistical metrics such as the mean and standard deviation. BTC has the most extended dataset, starting from January 1, 2017, with a mean price of 21,908.94 and a standard deviation of 18,749.33, indicating high volatility. ETH and BNB also have substantial datasets starting November 9, 2017, with mean prices of 1,381.28 and 190.99, respectively. Newer coins like SOL and AVAX have fewer data points, reflecting their recent launches. Coins with low mean prices, like DOGE, SHIB, and TRX, show smaller standard deviations, suggesting relatively lower volatility compared to high-value coins like BTC and BCH.

Table 1 Descriptive statistics of top cryptocurrencies

S/N	Cryptocurrency	Coins	Samples	Start date (dd/mm/yyyy)	End date (dd/mm/yyyy)	Mean	Std. Dev
1	BTC	BTC	2738	01/01/2017	30/06/2024	21,908.94	18,749.32
2	Ethereum	ETH	2426	09/11/2017	30/06/2024	1381.28	1195.18
3	Binance coin	BNB	2426	09/11/2017	30/06/2024	190.99	191.57
4	Solana	SOL	1543	10/04/2020	30/06/2024	56.29	60.04
5	Ripple	XRP	2426	09/11/2017	30/06/2024	0.52	0.32
6	Toncoin	TON	1039	27/08/2021	30/06/2024	2.35	1.50
7	Dogecoin	DOGE	2426	09/11/2017	30/06/2024	0.06	0.08
8	Cardano	ADA	2426	09/11/2017	30/06/2024	0.47	0.55
9	Tron	TRX	2426	09/11/2017	30/06/2024	0.05	0.03
10	Avalanche	AVAX	1380	13/07/2020	30/06/2024	31.50	26.63
11	Shiba Inu	SHIB	1171	17/04/2021	30/06/2024	0.00002	0.00001
12	Polkadot	DOT	1411	20/08/2020	30/06/2024	13.35	11.49
13	Chainlink	LINK	2426	09/11/2017	30/06/2024	9.46	9.44
14	BTC cash	BCH	2426	09/11/2017	30/06/2024	427.86	409.18
15	Unus sed leo	LEO	1868	21/05/2019	30/06/2024	3.06	1.64
16	NEAR protocol	NEAR	1356	14/10/2020	30/06/2024	4.64	3.82

The cryptocurrency market is highly interconnected, particularly during critical events, and its network structure evolves over time, providing new insights for investors aiming to optimize their portfolios and mitigate risks in the volatile cryptocurrency landscape [25]. While existing studies have been limited to mainly considering four popular coins, BTC, ETH, LTC, and XRP in their studies, few studies, such as the work of Akyildirim et al. [2] and Oyewola et al. [45] considering multiple cryptocurrencies, 12 and 15, respectively. To examine the intercorrelation among the 16 selected top coins and understand their correlation dynamics, Pearson correlation coefficients (PCC) were computed for all the coins using a heatmap, as depicted in Fig. 3.

The heatmap in Fig. 3 illustrates the PPC among the 16 selected cryptocurrencies, highlighting their interconnectedness within the market. To ensure uniformity in the analysis, daily closing price data from January 1, 2023, to June 30, 2024, was collected for all 16 coins, considering that each cryptocurrency has a different initial launch date. This uniform time frame allows for a fair comparison of correlations across all selected assets. BTC exhibits moderate to strong positive correlations with many other cryptocurrencies, with correlation coefficients above 0.7. This interconnected behaviour suggests that BTC often moves in tandem with other top coins, making it an ideal candidate to train and test the robustness and predictive power of the proposed model, Helformer. By focusing on BTC for initial model implementation, its market influence and significant correlation with other cryptocurrencies can be leveraged, ensuring that any insights or patterns identified are likely relevant to the broader cryptocurrency market.

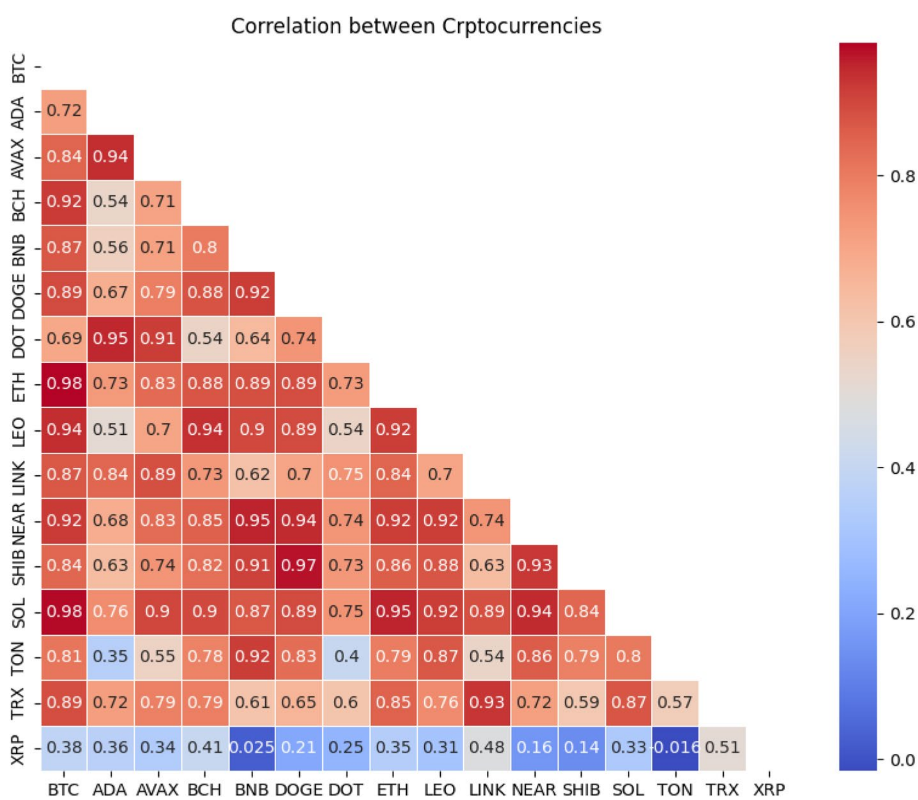

Fig. 3 Correlation heatmap

Additionally, it is essential to note that all the selected cryptocurrencies exhibit positive correlations with one another, indicating that their price movements tend to follow similar trends within the market.

Data preprocessing

First, the daily closing price data of BTC was downloaded from Yahoo Finance for the period between January 1, 2017, and June 30, 2024. Then, an exploratory data analysis was conducted to identify potential issues and ensure data quality. Upon examination, insights show that the data is of high quality with no missing values. A quick overview of the BTC dataset reveals that there are 2,738 observations recorded, with a minimum price of 777.75 USD and a maximum price of 73,083.50 USD within the given period. The mean price across all samples is 21,908.94 USD, and the standard deviation is 18,749.32 USD. Afterward, outliers are retained in the dataset as they provide significant information, particularly in the highly volatile cryptocurrency market, where extreme price fluctuations are common. This approach aligns with common practices in existing studies, where outliers are often preserved to reflect real-world conditions [67]. However, several strategies were employed to prevent the risk of overfitting while maintaining the model's predictive power. First, MinMax scaling was applied to normalize the data and prevent extreme values from dominating the learning process. Additionally, dropout layers were incorporated to reduce the model's sensitivity to outliers, while Bayesian hyperparameter tuning helped optimize model performance and avoid excessive fitting to noise. Following this, the dataset was cleaned to ensure there were no NaN values, further maintaining the integrity of the data in the current study.

Next, the BTC dataset was subjected to seasonality and stationarity tests. To achieve this, we utilized the "statsmodels" library in Python to perform a seasonal decomposition of the time series data. This decomposition allowed us to break down the data into its observed, trend, seasonal, and residual components, providing a clear visualization of underlying patterns and variations in the dataset. By analyzing these plots, we can better understand the cyclical behaviour and trends in BTC prices, which is crucial for building robust forecasting models. The seasonal decomposition plot in Fig. 4 breaks down the time series into four components: observed, trend, seasonal, and residual. The observed plot represents the original BTC price data from 01 January 2017 to 30 June 2024, showing significant volatility with notable peaks around 2021 and 2022, followed by periods of correction and recovery. The overall trend indicates an upward movement from 2017 to early 2021, followed by a decline until mid-2023, after which the trend rises again towards 2024. This long-term trend component smooths out short-term fluctuations, capturing the general direction of BTC prices, which suggests a potential for recovery or growth in the market after a significant decline.

The seasonal component illustrates repeating cyclical patterns throughout the yearly period of 365 days, indicating some level of periodicity in BTC price movements. These cycles could be driven by factors such as investor sentiment, market psychology, macroeconomic conditions, pandemics, or regular events like regulatory news or technological updates. The residual component captures the random noise and irregular fluctuations that are not explained by the trend or seasonal components. The residuals show significant volatility, particularly during periods of intense market activity like 2017–2018 and

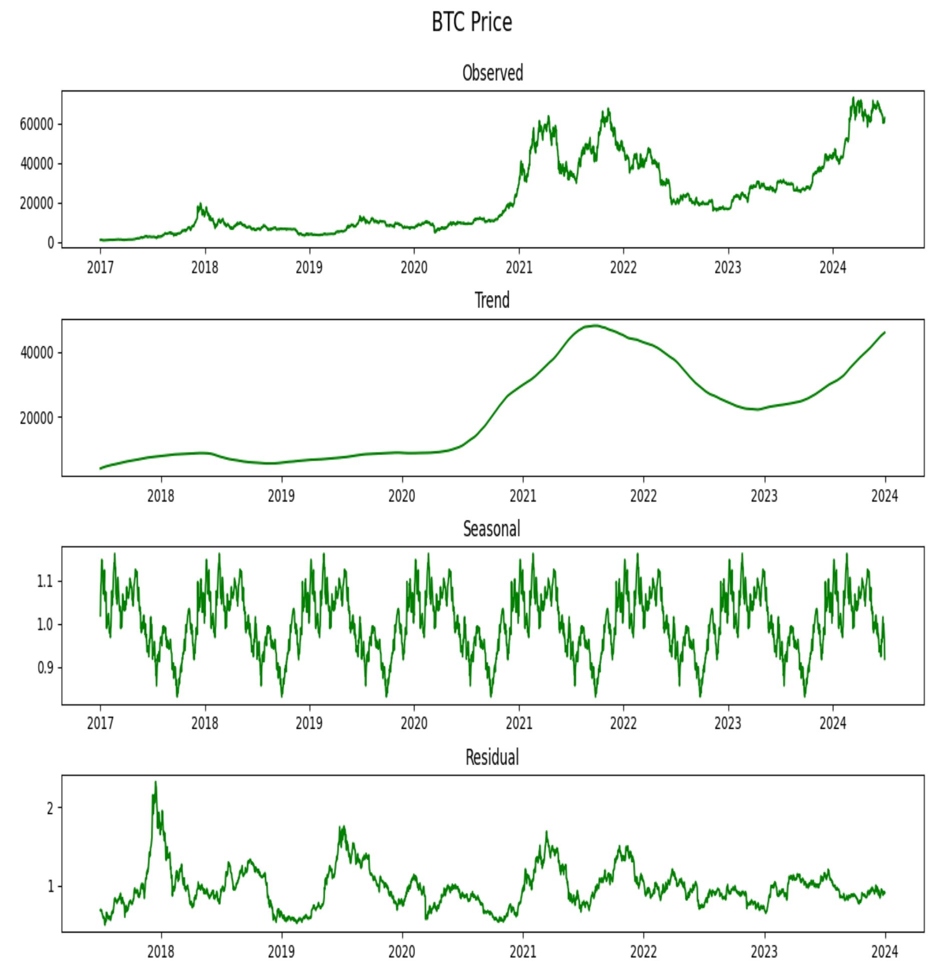

Fig. 4 Seasonal-trend decomposition plot—BTC

2021–2022, suggesting that there are unpredictable market shocks or events impacting BTC prices. This decomposition provides valuable insights for the proposed model to identify and separate predictable cyclical patterns from random, unforeseen variations, enabling a more subtle approach to predicting BTC price movements.

To further substantiate the claim regarding the seasonality and non-stationarity nature of cryptocurrency, an Autocorrelation Function (ACF) test was conducted, as shown in Fig. 5. The ACF plot measures the correlation between the time series data and its lagged values over different periods. From the ACF plot of the BTC closing prices, it is evident that there is a high level of autocorrelation at multiple lags, which gradually declines but remains significantly positive even after 50 lags. This persistent autocorrelation indicates that the BTC price series exhibits strong temporal dependencies and long-term memory effects. Such prolonged correlations confirm that the BTC price data is non-stationary, as the correlations do not diminish quickly to zero. This behaviour is typical for financial time series data, where past prices considerably impact future prices. The high autocorrelation across many lags supports the need for more sophisticated models like Helformer, which can effectively capture these long-range dependencies and provide more accurate forecasts.

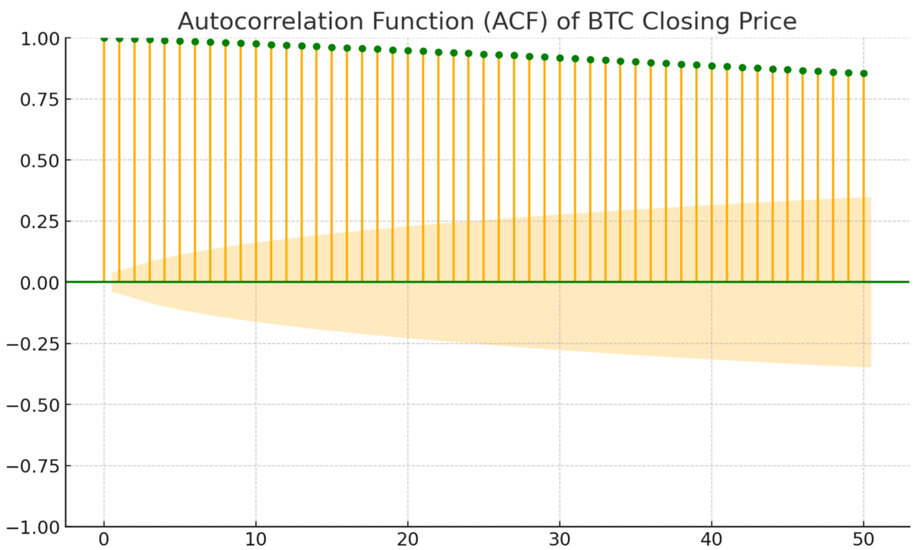

Fig. 5 ACF plot of BTC data

The non-stationarity further supports the need for sophisticated models like Helformer to effectively capture the complex patterns and temporal dependencies in BTC prices for robust prediction.

Experimental set-up

After preprocessing the data for model implementation, the proposed model will be implemented alongside five other models: RNN, LSTM, BiLSTM, GRU, and Transformer. The dataset is split into training and testing sets (80:20) to ensure a robust evaluation of each model's performance. Additionally, a validation split is set to 0.2. This validation step helps to fine-tune the models and prevent overfitting. The parameters used in the initial training phase are detailed in Table 2.

A time step of 30 was chosen because this window size has demonstrated better accuracy in previous studies, such as those by Dutta et al. [15], Chowdhury et al. [11], and Jin and Li [28]. The loss function was set at "mean square error," while the activation function was set at "Mish." The Mish activation function, a state-of-the-art activation function, is defined by the formula presented in Eq. 6.

$$f(x) = x \cdot tanh(ln(1 + ex)) \tag{6}$$

where $\ln(1 + e^x)$ is the softplus activation function.

This smooth, non-monotonic Mish function integrates a self-gating property, similar to the Swish function, allowing each neuron to adjust its output based on the input it receives. The smoothness of "Mish" ensures continuous derivatives, which are crucial for maintaining a steady gradient flow through deep networks. This can be particularly advantageous in preventing issues like gradient discontinuities during the learning process. Mish offers several benefits over traditional activation functions such as ReLU and Swish, particularly in its ability to mitigate the "dying ReLU problem" by avoiding zero-gradient regions [39]. Unlike ReLU, Mish allows for the propagation of negative values, which helps capture more

Table 2 Model setup parameters

Models	Helformer	Transformer	RNN/LSTM/BiLSTM/GRU
num_transformer_blocks	1	1	–
num_heads	4	4	–
head_size	16	16	–
dropout	0.1	0.1	0.1
epochs	100	100	100
batch_size	32	32	32
neurons	30	–	30
hidden_layers	–	–	1
learning_rate	0.001	0.001	0.001
optimizer	Adams	Adams	Adams
loss	MSE	MSE	MSE
ff_dim	–	16	–
activation function	Mish	Mish	Mish

complex patterns within the data. While tanh also handles negative values and offers a smooth gradient, it can lead to vanishing gradients in deeper networks, a limitation less pronounced in Mish due to its characteristics. These properties make Mish a promising choice for complex neural network tasks, including time series modelling, where understanding deep temporal dependencies is essential. The versatility of Mish as an activation function, surpassing ReLU and Swish, is demonstrated in the multiple experiments conducted by Sbrana and Lima de Castro [52]. Their study shows that neural network models with Mish activation function consistently generate lower prediction errors than their alternatives. Figure 6 provides a holistic framework for the entire model implementation and training.

Given that the data preprocessing phase is critical for the success of this experiment in accurately predicting cryptocurrency prices, BTC data were scaled to reduce noise and variability, thereby enhancing the model's ability to recognize underlying trends. This transformation is particularly important for stabilizing variance across the dataset, ensuring that price patterns remain distinct and interpretable for effective forecasting. To achieve this, MinMaxScaler is adopted, which normalizes values within a fixed range of 0 to 1, as shown in Eq. 7. The choice of MinMax scaling is based on its ability to preserve the relative relationships and distribution of the data while preventing extreme price fluctuations from dominating the learning process. Unlike standardization methods such as Z-score normalization, which assumes a Gaussian distribution and centers data around a mean of zero, MinMax scaling retains the original structure of the data, making it more suitable for highly volatile financial time series. Additionally, this scaling technique helps mitigate vanishing or exploding gradient issues in deep learning models by ensuring that input values remain within a constrained range, improving convergence efficiency during training.

$$\bar{y}_t = \frac{y_t - \min(y_t)}{\max(y_t) - \min(y_t)} \tag{7}$$

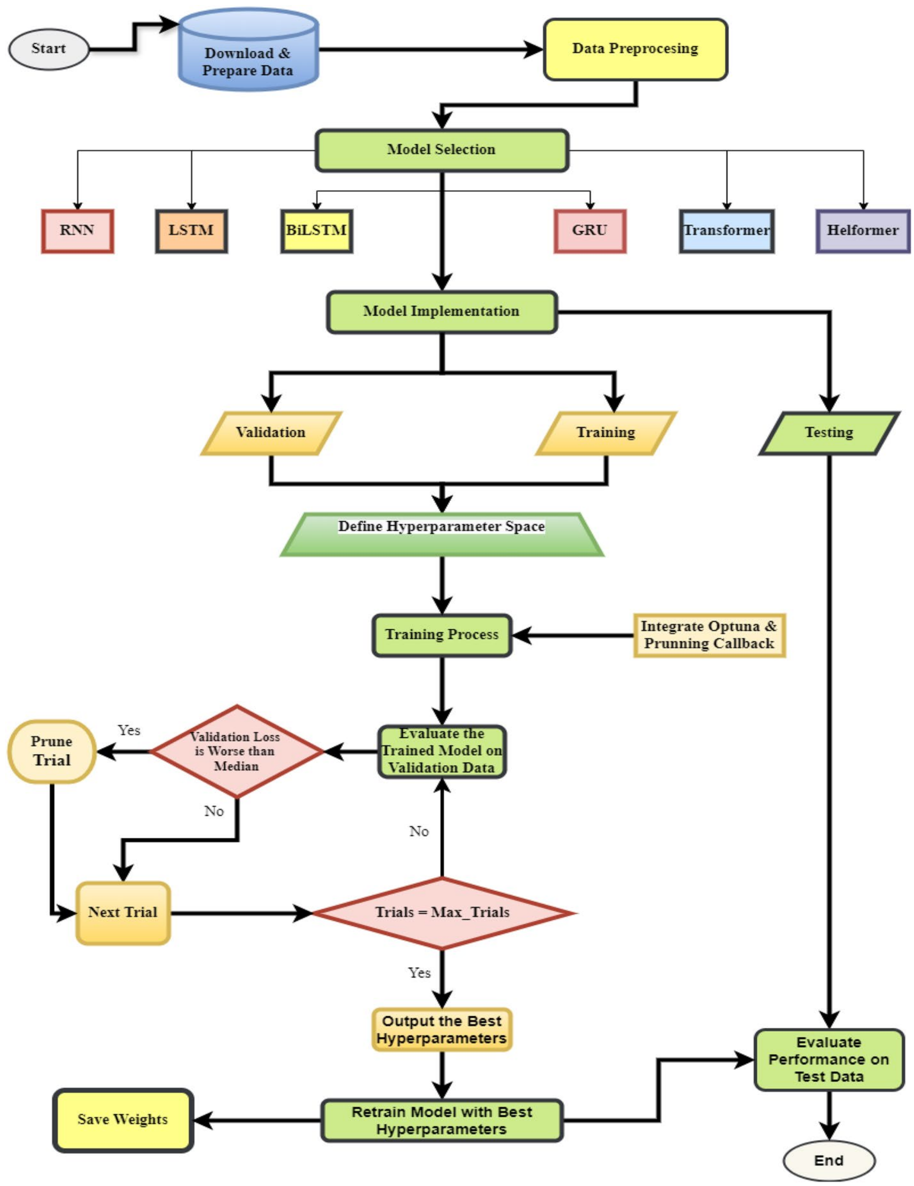

Fig. 6 Systematic framework

where $\bar{y}_t$ denotes the normalized price at any time t, while y_t is the smoothed price at any time t.

For the execution of all models in this study, Python 3.10.12 was utilized on Google Colab, a choice driven by the platform's capacity to provide efficient and accessible computing resources. Google Colab offers a user-friendly environment that supports intensive computational tasks by providing access to external hardware accelerators and compute units. This significantly reduces the computational load, making it ideal for handling the robust needs of deep learning models. The environment runs TensorFlow 2.17.0 and incorporates the Keras library, which comes pre-equipped with a wide array of deep learning models and libraries ready for use. Data processing and visualization tasks were primarily conducted using the Python libraries: Matplotlib and Seaborn.

Given the high computational demands of the proposed models, particularly during the hyperparameter tuning phase, the premium version of Google Colab was considered, which includes access to the NVIDIA A100 GPU. This advanced GPU enhances computing power, accelerates processing speed, and expands computational capabilities, which are crucial for managing the intense demands of predictive models. The A100 GPU is particularly valued for its high-performance computing abilities, making it an excellent tool for data-intensive tasks and ensuring efficient execution of deep learning frameworks.

Hyperparameters optimization process

Hyperparameter optimization is a crucial stage in machine learning training. It aims to optimize the parameters that control the learning process, resulting in the highest potential model performance. The selection of suitable hyperparameters is crucial as they have a direct impact on the training model, which learns from the data and makes accurate projections on unseen data. Inadequately selected hyperparameters can result in problems such as overfitting, underfitting, or ineffective learning, which eventually diminish the model's capacity to accurately forecast and its reliability. Three of the most commonly used tuning strategies are Grid search, Random search, and Bayesian search. Grid search is widely utilized due to its straightforward implementation and ease of parallelization, as well as its dependability in low-dimensional spaces and the reproducibility of tuning results. However, grid search faces significant challenges, particularly in high-dimensional spaces, where the number of trials grows exponentially with the increase in hyperparameters, a phenomenon often referred to as the curse of dimensionality [6].

In contrast, random search selects hyperparameters by drawing independent samples from a uniform distribution [6]. Random search retains many of the practical advantages of grid search, including simplicity and reproducibility, but offers a significant performance boost in high-dimensional hyperparameter spaces. Bayesian optimization takes a fundamentally different approach to hyperparameter tuning when compared with the others by building a surrogate model of the hyperparameter response function instead of exhaustively sampling the hyperparameter space [49]. It uses this surrogate model to inform the search process and selects explicitly the next set of hyperparameters to evaluate and reduce the uncertainty of the model. The running of the machine learning model is then assessed with these hyperparameters, updating the probabilistic model and creating a posterior distribution that guides future selections. This iterative process continues until improvements are minimal or computational resources are exhausted, ultimately yielding the optimal hyperparameter configuration. Bayesian search is particularly efficient, often requiring fewer evaluations to locate the optimal solution. Equation 8 is used to find the maximum value of the unknown objective function:

$$x^* = \arg\max_{x \in X} f(x) \tag{8}$$

Here, X represents the search space of hyperparameters, denoted by x.

In Bayesian optimization, the objective function f is treated as a random function, and a prior distribution is assumed over it. This optimization approach hinges on two crucial elements: the prior function and the posterior function, the latter typically represented by an acquisition function. The prior function models the expected behaviour of the objective function and is often estimated using methods such as Gaussian Processes (GP) or more specialized algorithms like the Tree-structured Parzen Estimator (TPE) [22]. As evaluations of the function are collected, the prior is updated to form a posterior distribution, which captures insights from new data and refines the understanding of the function's behaviour. This posterior distribution is essential for constructing an acquisition function (u), which strategically guides the selection of the next query point for evaluation, aiming to optimize the search process. Common choices for the acquisition function include the Probability of Improvement (PI) and Expected Improvement (EI), both designed to steer the search towards regions of the hyperparameter space that promise the most significant enhancements. The PI function, in particular, focuses on exploring areas around the current optimal point to find potentially superior values. This exploration is crucial for efficiently navigating the search space and is formalized in Eq. 9, which calculates the probability that a new sample will yield an improvement over the current best observation.

$$PI(x) = \varphi(\frac{\mu(x) - f(x^{+})}{\sigma(x)}) \tag{9}$$

In this context, ϕ represents the cumulative distribution function (CDF) of the Gaussian distribution.

The PI acquisition function in Bayesian optimization has a key limitation: it tends to focus sampling efforts near the current optimal solution, emphasizing exploration. This can lead to potentially better solutions being overlooked if they lie farther from the localized optimum, potentially causing the model to get stuck in local optima. To mitigate this issue, the EI acquisition function is often utilized. The EI function systematically explores the vicinity of the current optimum and calculates the expected improvement for each new point evaluated. If the calculated EI at a new point falls below a predetermined threshold, it is inferred that the current optimal point is likely the best solution within that region. Consequently, the algorithm then shifts its focus to explore other areas of the search domain, thus effectively balancing exploration with exploitation. This balance is crucial for avoiding local optima and ensuring a more comprehensive search of the hyperparameter space. The degree of improvement (I), which is the difference between the function value at the newly selected point and the value at the current optimum, is central to this process [22]. Suppose the new point's function value does not surpass the current optimal value. In that case, the improvement is considered zero, as depicted in Eq. 10. This mechanism ensures that the optimization process continuously moves towards discovering potentially superior solutions.

$$I(x) = \max\left\{0, f_{t+1}(x) - f(x^{+})\right\} \tag{10}$$

Equation 11 and Eq. 12 represent the probability density function for I and EI.

$$f(I) = \frac{1}{\sqrt{2\pi}\sigma(x)} \exp(-\frac{(\mu(x) - f(x^{+}) - I)^2}{2\sigma^2(x)}), I \geq 0 \tag{11}$$

$$EI = \sigma(x)[Z\varphi(Z) + \varphi(Z)] \quad (12)$$

where ϕ is the probability distribution function of the standard normal distribution Z in Eq. 13.

$$Z = \frac{\mu(x) - f(x^{+})}{\sigma(x)} \quad (13)$$

In this work, Bayesian optimization was employed to fine-tune the hyperparameters of the Helformer model and other deep learning baselines (RNN, LSTM, BiLSTM, GRU, and Transformer). Unlike grid or random search, Bayesian optimization efficiently explores the search space using a probabilistic surrogate model, reducing the number of function evaluations needed to find the optimal hyperparameters. This study utilized TPE algorithm from the Optuna framework, which models the objective function as a probabilistic distribution and selects hyperparameter values that maximize EI. The optimization process follows these key steps:

1. Define the Search Space: This is achieved by specifying the possible values for each hyperparameter (e.g., learning rate, dropout rate, batch size).
2. Initialize Random Trials: The algorithm first evaluates a few randomly chosen configurations to build an initial model.
3. Build a Surrogate Model: A probabilistic model is constructed to approximate the objective function.
4. Select the Next Set of Hyperparameters: Based on the EI criterion, the next promising hyperparameters are selected.
5. Evaluate and Update the Model: The new hyperparameter combination is tested, and the surrogate model is updated iteratively.
6. Convergence: The process stops when performance gains become negligible or when a set number of trials is reached.

To ensure efficiency, the number of trials is set to 50, and the Optuna Pruner feature is enabled to terminate underperforming trials early, preventing unnecessary computations. The optimization direction is set to minimize the MSE as the primary objective. The search space for each model is detailed in Table 3, specifying the hyperparameter ranges explored during Bayesian optimization.

Evaluation metrics

Six evaluation metrics were employed to assess the predictive prowess of the developed models, and they were categorized into similarity-based and dissimilarity-based metrics. The similarity-based metrics include R-squared (R^2), Explained Variance Score (EVS), and Kling-Gupta Efficiency (KGE). R^2 measures the proportion of the variance in the dependent variable that is predictable from the independent variables, indicating the goodness of fit of the model. EVS assesses the proportion of the variance in the target variable accounted for by the model, reflecting the model's capability to explain data variability. KGE combines the

Table 3 Bayesian optimization search space

Hyperparameters	RNN/LSTM/BiLSTM/GRU	Transformer	Helformer
neurons	[20, 50] (step = 5)	–	[20, 50] (step = 5)
layers	[1, 2]	–	–
num_blocks	–	[1, 4]	[1, 4]
learning_rate	[0.0001, 0.01]	[0.0001, 0.01]	[0.0001, 0.01]
dropout_rate	[0, 0.3]	[0, 0.3]	[0, 0.3]
batch_size	[16, 32, 64, 128]	[16, 32, 64, 128]	[16, 32, 64, 128]
epochs	[50, 150] (step = 5)	[50, 150] (step = 5)	[50, 150] (step = 5)
num_heads	–	[2, 10] (step = 2)	[2, 10] (step = 2)
head_size	–	[8, 64] (step = 8)	[8, 64] (step = 8)
ff_dim	–	[16, 64] (step = 16)	–

Pearson correlation coefficient, bias ratio, and variability ratio to provide a balanced measure of correlation, bias, and variability error between observed and predicted values.

On the other hand, the dissimilarity-based metrics include Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE). Together, these metrics comprehensively evaluate each model's performance, capturing both the alignment and deviation between predicted and actual values. Equation 14–19 represents the formulas for the six evaluation metrics used to assess the performance of the developed models. These metrics provide a comprehensive understanding of both the similarity and dissimilarity between the predicted and actual values.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{x}_i - x_i)^2} \tag{14}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\overline{x}_i - x_i}{x_i}\right| * 100\% \tag{15}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\overline{x}_i - x_i| \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i-1}^{N}(\overline{x}_i - x_i)^2}{\sum_{i-1}^{N}(\hat{x} - x_i)^2} \tag{17}$$

where, x_i are the actual values, $\overline{x}_i$ are the predicted values, $\hat{x}$ is the mean of the actual values, and N is the length of the dataset.

$$EVS = 1 - \frac{Var(x - \overline{x})}{Var(x)} \tag{18}$$

where $Var(x)$ denotes the variance of the actual values and $Var(x - \overline{x})$ is the variance of the errors.

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \quad (19)$$

where r is the Pearson correlation coefficient, α is the variability ratio, and β is the bias ratio.

Empirical results and discussions

This section presents the results and discussion of the base models used in this study. After applying hyperparameter tuning using Optuna, based on the hyperparameter space outlined in Table 3, optimized parameters were obtained for training the final version of the Helformer model alongside five other sophisticated models: RNN, LSTM, BiLSTM, GRU, and Transformer. The results of these optimized models are presented and discussed, showcasing significant improvements in predictive performance due to the fine-tuning process. Furthermore, a trading strategy was implemented to demonstrate the practical applicability of each model by comparing their performance to the traditional B&H strategy. The results from these trading strategies provide insight into the potential financial gains and risk management capabilities of the individual models. To further validate the versatility and robustness of the proposed model, a comparative analysis was conducted by replicating the experimental setups and parameters from notable works in the literature, using their datasets to benchmark the performance of the proposed model against existing models. Lastly, this section highlights the cross-learning ability of the Helformer model, which was initially trained using BTC data. The saved weights from this pre-trained model were then applied to 15 other top cryptocurrencies. This approach demonstrated the model's exceptional predictive accuracy and significant returns when employed in trading strategies across different cryptocurrencies, highlighting the model's generalizability and effectiveness in diverse market conditions.

Results of the base models

This study applied the experimental setup described earlier to build all the selected models using their default configurations without hyperparameter tuning. The initial results provide an overview of evaluation metrics, including RMSE, MAPE, MAE, R^2, EVS, and KGE on the test data. Table 4 presents the performance of the base models before any hyperparameter tuning, revealing significant differences in their predictive accuracy. Among the models, the Helformer stands out with exceptional performance across all

Table 4 BTC base model – Evaluation metrics on test data

Model	RMSE	MAPE	MAE	R^2	EVS	KGE
RNN	1256.3767	2.3942%	915.7597	0.9941	0.9952	0.9851
LSTM	1426.5453	3.1121%	1123.4248	0.9924	0.9930	0.9669
BiLSTM	1331.3047	2.6030%	980.5543	0.9933	0.9937	0.9862
GRU	1314.9097	1.9241%	830.1504	0.9935	0.9944	0.9674
Transformer	1657.1426	3.0053%	1174.7753	0.9897	0.9900	0.9855
Helformer	16.0822	0.0343%	13.4487	1	1	0.9995

evaluation metrics. The Helformer achieves the lowest RMSE (16.0822). Its MAPE is also impressively low at 0.0343%, showcasing superior accuracy compared to the other models. The MAE for the Helformer is 13.4487, further highlighting its precision in prediction. The Helformer model also achieves perfect scores for R^2 and EVS (both equal to 1), indicating that it perfectly captured the variance in BTC prices. The high KGE score of 0.9995 shows a nearly perfect agreement between the observed and predicted values. The RNN model, which is a simpler recurrent neural network architecture, shows a significantly higher RMSE of 1256.3767 and MAPE of 2.3942%. The MAE is also high at 915.7597, indicating that the model has a relatively large average error in predictions. Although the R^2 value of 0.9941 and EVS of 0.9952 are still high, suggesting a good fit to the data, the model's errors indicate room for improvement. The LSTM model, known for its capability to manage long-term dependencies in time series data, records an RMSE of 1426.5453, MAPE of 3.1121%, and MAE of 1123.4248. These results suggest that, although LSTM is an effective model for time series forecasting, it underperforms compared to the Helformer. The lower R^2 (0.9924) and EVS (0.9930) compared to the Helformer indicate that LSTM does not capture the variance in BTC prices as well. The BiLSTM model, a more advanced version of LSTM that captures dependencies in both forward and backward directions, shows some improvement over LSTM with an RMSE of 1331.3047 and MAPE of 2.6030%. However, its MAE of 980.5543 and slightly lower R^2 (0.9933) compared to the Helformer indicate it still lacks the precision and robustness needed for optimal forecasting.

The GRU model performs slightly better than the LSTM and BiLSTM models with an RMSE of 1314.9097 and a lower MAPE of 1.9241%. The MAE of 830.1504 is also lower than that of the LSTM and BiLSTM. However, the R^2 (0.9935) and EVS (0.9944) are still below those achieved by the Helformer, indicating that while GRU is effective, it does not perform as well as the Helformer. The Transformer model, which utilizes self-attention mechanisms, records the highest RMSE (1657.1426) and a relatively high MAPE of 3.0053%. The MAE is also the highest among the models at 1174.7753, indicating substantial prediction errors. Despite having a high R^2 value of 0.9897, the Transformer model's performance in this context is not as efficient as the Helformer. In sum, the Helformer clearly outperforms all other models in their base configurations, demonstrating superior prediction accuracy and robustness. Its outstanding performance across all metrics suggests that its architecture, which incorporates series decomposition and attention mechanisms, is particularly well-suited for handling the complex and volatile nature of cryptocurrency data.

Results of the optimized models

The optimal hyperparameter values for each model, obtained through Bayesian optimization, are as follows: For the Transformer model, the optimal configuration includes a feed-forward dimension of 16, 2 blocks, a learning rate of 0.0085, a dropout rate of 0.0181, batch size of 16, and 100 epochs. Additionally, it utilizes 10 attention heads with a head size of 32. The RNN, LSTM, BiLSTM, and GRU models were optimized with unit sizes of 40, 45, 40, and 40, respectively, with layers set at 2, 1, 1, and 1. Their learning rates were tuned to 0.0058, 0.0084, 0.0087, and 0.0082, while dropout rate were 0.0117, 0.1685, 0.0321, and 0.0001, respectively. Batch sizes varied as 64, 16, 16, and 64, with the number of training epochs optimized at 85, 130, 130, and 85, respectively. The Helformer model, which demonstrated superior performance, was optimized with 20 units, 1 block, a learning rate of 0.0037, a dropout rate of 0.0194, and a batch size of 16, trained for 95 epochs. The model was configured with 4 attention heads and a head size of 48. Table 5 presents the optimized results of the models after hyperparameter tuning, demonstrating their improved performance in predicting BTC prices on the test dataset. It reveals that the Helformer model, after optimization, significantly outperforms all other models across all evaluation metrics. The Helformer achieves an exceptionally low RMSE of 7.7534, indicating that the deviation between its predicted and actual BTC prices is exceptionally minimal. The MAPE is remarkably low at 0.0148%, showcasing its outstanding accuracy in predicting BTC prices. The MAE is also the lowest among all models at 5.9252, demonstrating high precision. The R^2 and EVS metrics both equal 1, signifying that the Helformer model perfectly explains the variance in BTC prices, indicating a perfect fit. The KGE of 0.9998 suggests near-perfect agreement between observed and predicted values, further validating its effectiveness in capturing the complex dynamics of BTC prices.

Comparatively, the other models: RNN, LSTM, BiLSTM, GRU, and Transformer, also show improved performance after hyperparameter tuning but still fall short of the Helformer in terms of accuracy and precision. The BiLSTM model, for example, achieves an RMSE of 1140.4627 and MAPE of 1.9514%, which are substantial improvements compared to its base model performance. However, its MAE of 766.7234 and R^2 of 0.9951 indicate that it still has larger errors and slightly less explanatory power compared to the Helformer. The RNN model also shows good performance with an RMSE of 1153.1877, MAPE of 1.9122%, and MAE of 765.7482. Its R^2 value of 0.9950 and EVS of 0.9951 are both high, suggesting that the model fits the data well. However, the prediction errors are larger than those of the Helformer. The GRU model performs similarly to the RNN, with an RMSE of 1151.1653, MAPE of 1.7500%, and MAE of 724.5279. Although it demonstrates slightly better performance than RNN, with a lower MAPE and MAE, its overall accuracy and precision are still inferior to those of the Helformer. Also, The LSTM model records an RMSE of 1171.6701, MAPE of 1.7681%, and MAE of 737.1088, reflecting improvements from its base performance but still lagging behind in comparison to the Helformer. The Transformer model, while known for its strong performance in sequence-to-sequence tasks, shows an RMSE of 1218.5600, MAPE of 1.9631%, and MAE of 799.6003. Despite its high R^2 (0.9944) and EVS (0.9946) values, the Transformer model has the highest prediction errors among the optimized models, suggesting it is

Table 5 BTC optimized model – Evaluation metrics on test data

Model	RMSE	MAPE	MAE	R^2	EVS	KGE
RNN	1153.1877	1.9122%	765.7482	0.9950	0.9951	0.9905
LSTM	1171.6701	1.7681%	737.1088	0.9948	0.9949	0.9815
BiLSTM	1140.4627	1.9514%	766.7234	0.9951	0.9952	0.9901
GRU	1151.1653	1.7500%	724.5279	0.9950	0.9950	0.9878
Transformer	1218.5600	1.9631%	799.6003	0.9944	0.9946	0.9902
Helformer	7.7534	0.0148%	5.9252	1	1	0.9998

less suitable for this particular time series forecasting task without further adjustments. The significant reduction in prediction errors and the perfect fit metrics (R^2 and EVS) for the Helformer model bring to light the effectiveness of its architecture and Optuna tuning process. This highlights the Helformer model's potential as a powerful tool for forecasting cryptocurrency prices in volatile markets.

Figure 7 illustrates the outstanding performance of the Helformer model, which exhibits a very accurate alignment with the true data, suggesting the most negligible error in predictions. The Helformer model demonstrates a remarkable level of precision, indicating its superior ability to capture the intricate dynamics of cryptocurrency data compared to the other models discussed. The Helformer model's precise fit demonstrates its usefulness and provides a reliable tool for investors, analysts and researchers seeking to make well-informed financial judgements. In sum, the empirical results justify the introduction of the series decomposition component, the attention mechanism, and the replacement of the FFN with an LSTM component in the proposed Helformer model. These components collectively enhance the model's ability to deal with the volatility, seasonality, non-stationarity, and non-linearity of time series data, leading to highly accurate predictions that are critical for effective cryptocurrency forecasting.

Implementation of trading strategy

This section discusses implementing a simple trading strategy to assess the practical applicability of the optimized models in generating financial returns from trading BTC. The results of this trading strategy are presented in Table 6 and Fig. 8, which provide key performance indicators such as Excess Return (ER), Volatility (V), Maximum Drawdown (MDD), and Sharpe Ratio (SR) for each model and the Buy & Hold (B&H) strategy.

A trading strategy is formulated using ER, V, MDD, and SR. If the forecasted value $\overline{x}_{t+1}$ for the next day exceeds the most recent observed value x_t, the strategy would initiate a

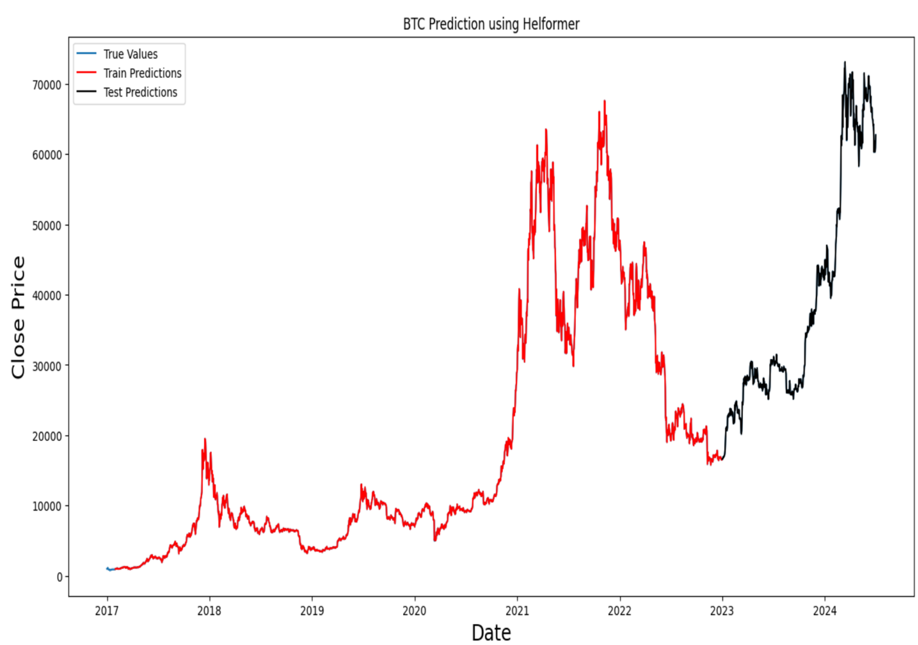

Fig. 7 BTC—predicted curves vs True curve

Table 6 Trading Strategy – BTC

Models	Excess Return (ER)	Volatility (V)	Max Drawdown (MDD)	Sharpe Ratio (SR)
RNN	157.57%	0.0246	−0.1871	2.2146
LSTM	90.88%	0.0247	−0.1617	1.2611
BiLSTM	171.23%	0.0246	−0.1507	2.4117
GRU	84.76%	0.0248	−0.2061	1.1743
Transformer	47.62%	0.0248	−0.4369	0.6488
Helformer	925.29%	0.0178	$-0.1943*10^{-4}$	18.0604
B&H	277.01%	0.0247	−0.1477	1.8529

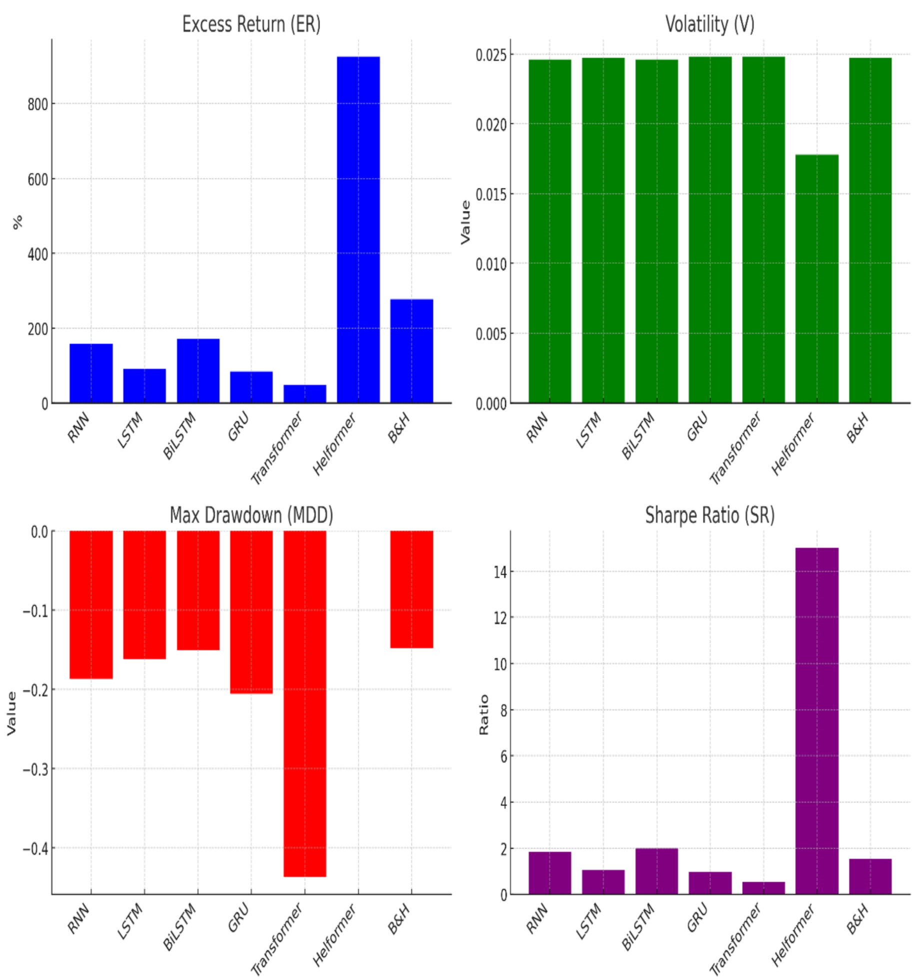

Fig. 8 Trading results

long one position in the index. Alternatively, if $\overline{x}_{t+1}$ is lesser than x_t, it would initiate a short one position index. Perhaps there is no difference; no position is held. The calculation of the return at any particular time $t + 1$ is determined according to Eq. 20:

$$R_{t+1} = \ln \frac{x_{t+1}}{x_t} * sign(\overline{x}_{t+1} - x_t) \tag{20}$$

The sign (.) represents the sign function, which returns + 1 if the argument is positive, -1 if negative, and 0 if zero. The net value (NV) of the strategy, which represents the total return, is calculated using Eq. 21, where $NV_1 = 1$ and $t > 1$. Also, since transaction costs vary across different exchanges and asset types, a 1% transaction cost is assumed to account for potential variations. For example, Binance charges 0.1% for spot trading, but fees may differ across platforms or for different cryptocurrencies.

$$NV_t = 1 + \sum_{i=2}^{t} R_t \tag{21}$$

Volatility is a term that quantifies the degree of change in the value of a security, index, or market across a given period. It plays a crucial role as a tool for investors and traders to evaluate risk and make well-informed decisions. Equation 22 is commonly used in computing volatility.

$$V = \sigma(R_t) \tag{22}$$

where σ represents the standard deviation of returns.

Maximum drawdown is a risk indicator that quantifies the most significant decline in the value of a portfolio or investment from its highest point to its lowest point before reaching a new high. It is frequently employed to assess the risk associated with a particular investment or compare various asset risk levels. Equation 23 is commonly used in computing maximum drawdown.

$$MDD = \max_{i<j} \frac{NV_j - NV_i}{NV_i} \tag{23}$$

The Sharpe Ratio is a financial metric that quantifies an investment's performance to its level of risk. The Sharpe ratio measures the additional return gained per unit of risk assumed in an investment. The Sharpe Ratio can be calculated using Eq. 24.

$$SR = \frac{R_t - R_f}{\sigma} \tag{24}$$

R_f represents risk free interest rate. In this stuudy, R_f is assumed to be 1%.

Table 6 illustrates the effectiveness of the different models in a trading context by showing their ability to maximize returns while minimizing risk. Among all models, the Helformer model stands out remarkably, achieving an ER of 925.29%. This return is significantly higher than that of any other model, indicating the Helformer's exceptional capability to generate profit in the volatile cryptocurrency market. Additionally, the Helformer demonstrates the lowest V of 0.0178, suggesting it maintains relatively stable performance. The MDD for Helformer is nearly negligible at -0.1943*10^{-4}, indicating minimal risk of substantial loss during the trading period. Its SR, which measures the risk-adjusted return, is extraordinarily high at 18.0604, confirming that the Helformer not only generates high returns but also does so with an excellent risk management profile. In comparison, the other models show significantly lower performance across all metrics. The BiLSTM model has the second-highest ER of 171.23% with a volatility of 0.0246, which is comparable to other models except Helformer. The MDD for BiLSTM is relatively low at -0.1507, and the SR is 2.0039, indicating a good balance of return and risk. However, its performance is still far behind that of the Helformer model.

The RNN model also performs relatively well, with an ER of 157.57% and a volatility of 0.0246. Its MDD is -0.1871, which shows moderate risk levels, and its SR of 1.8401 indicates good risk-adjusted returns. However, it is less effective than BiLSTM and significantly underperforms compared to Helformer. The LSTM model records an ER of 90.88%, a volatility of 0.0247, and an MDD of -0.1617. Its SR is 1.0479, which suggests that while it provides a positive return, it does so with relatively higher risk compared to RNN and BiLSTM. The GRU model performs slightly worse than LSTM, with an ER of 84.76% and an MDD of -0.2061. Its volatility is slightly higher at 0.0248, and it has the lowest SR among the models (excluding the Transformer) at 0.9757, suggesting it is less effective in providing risk-adjusted returns. The Transformer model shows the weakest performance, with an ER of 47.62%, the highest MDD of -0.4369, and an SR of 0.5391. This indicates that the model has difficulty maintaining stable performance in the highly volatile cryptocurrency market and generates low returns relative to the risk taken.

The Buy & Hold (B&H) strategy, a traditional investment approach, results in an ER of 277.01%, volatility of 0.0247, and an MDD of -0.1477. Its SR of 1.8529 suggests that while it performs better than most models except for Helformer and BiLSTM, it is still not as effective as the Helformer model in balancing returns and risks. In sum, the results in Table 6 and Fig. 8 clearly demonstrate that the Helformer model significantly outperforms all other models and the B&H strategy in terms of excess return, risk management, and risk-adjusted returns. Its ability to achieve such high returns with minimal volatility and drawdown highlights the robustness and effectiveness of the Helformer model for practical cryptocurrency trading strategies. This performance validates the model's superior predictive capabilities and its potential as a valuable tool for investors, analysts, and asset managers in the cryptocurrency market.

Figure 9 illustrates the Net Value curves of various models and B&H strategy for BTC over the period from January 2023 to June 2024. The Net Value curve is a crucial indicator of how well a trading strategy performs over time, showing the cumulative return of an initial investment as it evolves. From the plot, it is evident that the Helformer model (represented in black) significantly outperforms all other models and the B&H strategy in terms of net value growth. The Helformer curve shows a steady, upward trajectory throughout the period, indicating its robust and consistent performance in generating returns from BTC trading. Unlike the other models and the B&H strategy, Helformer shows an almost exponential growth pattern, with a rapid increase in net value beginning around mid-2023. This suggests that the model effectively captures market trends and executes profitable trades, leading to substantial gains. In contrast, the net value curves of the other models. RNN, LSTM, BiLSTM, GRU, and Transformer are relatively flat, with modest upward trends. The BiLSTM model (cyan) shows a better performance than the RNN (pink), LSTM (green), GRU (blue), and Transformer (orange), indicating some capacity to capture and profit from market movements. However, the growth is much slower and less pronounced compared to Helformer. The RNN and LSTM models perform similarly, showing slight upward trends, but their curves are still much lower than that of Helformer, indicating lower profitability. While having some upward movement, the GRU and Transformer models remain the least effective, with the Transformer

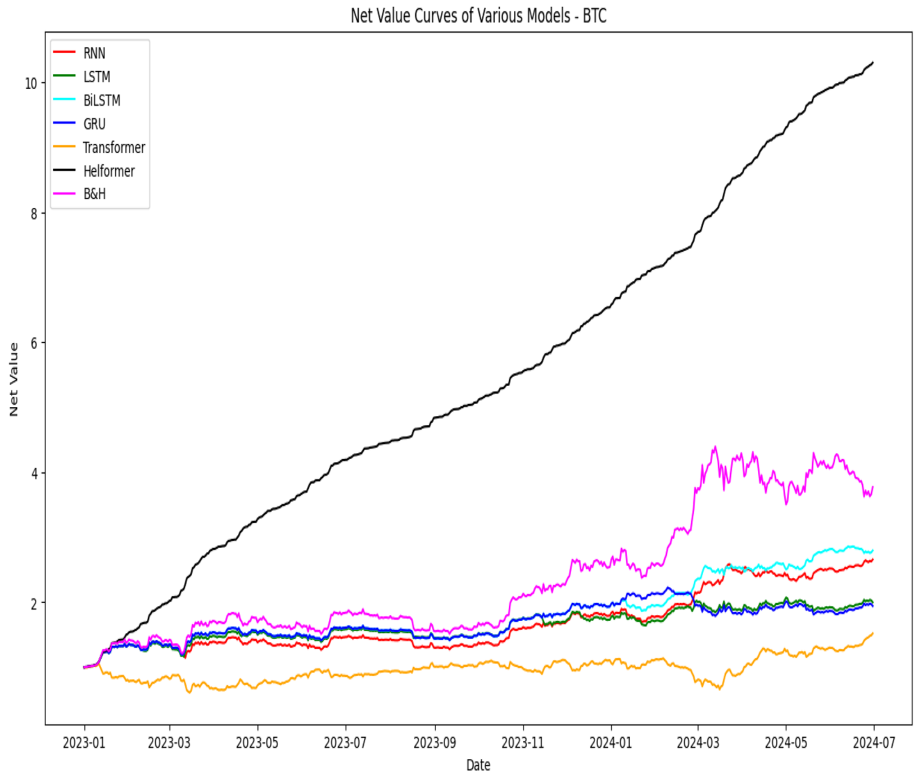

Fig. 9 BTC—net value curves

model, in particular, showing the flattest curve and the least net value growth, underscoring its limitations in this context.

The B&H strategy (purple) shows a stable but relatively moderate increase in net value, outperforming most models except Helformer. This demonstrates that while B&H is a safer strategy compared to some deep learning models, it does not capitalize on short-term market opportunities as effectively as Helformer does. In sum, the Net Value curves highlight the superior performance of the Helformer model in the context of BTC trading. Its ability to achieve continuous and substantial net value growth without significant drawdowns underscores its effectiveness in generating high returns with a robust risk management strategy. The other models, while offering some value, do not come close to matching Helformer's performance, reinforcing its status as the most suitable model for profitable cryptocurrency trading.

Comparison of helformer with existing studies

To showcase the versatility and robustness of the Helformer model, this study compares its performance with those reported in the latest and notable existing studies on cryptocurrency price prediction, specifically those using BTC as the prediction object. The comparison primarily focuses on evaluating the predictive accuracy of the Helformer model against a range of models from recent studies. This involved utilizing an identical dataset, applying the same data preprocessing techniques, and adopting similar data splitting strategies to ensure a fair and rigorous comparative analysis. Additionally, this study maintained consistent experimental setups and parameters as outlined in the selected studies to provide a direct and unbiased comparison. The chosen studies for this comparative analysis include a variety of models: singular models, hybrid models, and ensemble models, representing some of the most effective approaches in recent cryptocurrency research. These notable works include Hansun et al. [23], Seabe et al. [53], Jin and Li [28], and Fallah et al. [16], which have employed various state-of-the-art techniques to enhance prediction accuracy and trading strategies. By benchmarking Helformer against these diverse and advanced methodologies, this study aims to highlight its superior capabilities in terms of prediction accuracy, robustness across different market conditions, and generalization ability across multiple cryptocurrencies. This comprehensive comparison presented in Table 7 strengthens Helformer's position as a versatile and reliable model for cryptocurrency price forecasting, capable of outperforming both traditional and cutting-edge models presented in the current literature.

Table 7 Comparison of the Helformer model with existing studies

S/N	Models	RMSE	MAPE	MAE
Fallah et al. [16]				
1	ARIMA	13,178.34	38.20%	11,654.64
2	SVR	1043.95	3.000%	818.47
3	RF	1038.08	3.00%	731.72
4	DNN	784.42	2.10%	588.16
5	DNN + VAR	711.40	1.80%	508.49
6	Helformer	36.23	0.10%	27.86
Jin and Li [28]				
1	ARIMA	253.051	1.61%	172.681
2	RF	372.773	2.78%	283.246
3	SVM	330.389	2.23%	236.284
4	Informer	333.124	2.48%	257.918
5	Autoformer	402.196	3.08%	319.257
6	LSTM	275.958	1.82%	193.817
7	GRU	260.502	1.69%	180.501
8	EMD-AGRU-LSTM	223.556	1.75%	181.721
9	VMD-AGRU-GRU	150.032	1.04%	113.32
10	VMD-GRU-LSTM	127.284	0.88%	94.895
11	VMD-AGRU-LSTM	124.657	0.87%	93.756
12	VMD-AGRU-RESEMD-LSTM	105.13	0.75%	80.417
13	VMD-AGRU-RESVMD-LSTM	50.651	0.39%	42.298
14	Helformer	0.201	0.0014%	0.153
Seabe et al. [53]				
1	LSTM	1031.3401	3.94%	-
2	BiLSTM	1029.3617	3.56%	-
3	GRU	1274.1706	5.72%	-
4	Helformer	19.7973	0.050%	
Hansun et al. [23]				
1	LSTM	2518.0217	4.218%	1617.7592
2	BiLSTM	2222.7354	3.800%	1422.1933
3	GRU	1777.306	3.492%	1167.3461
4	Helformer	8.0665	0.010%	3.7670

Table 7 provides a comprehensive comparison of the Helformer model against various models reported in recent studies. Compared to Fallah et al. [16], where models like ARIMA, SVR, RF, DNN, and DNN + VAR show higher RMSE (from 711.40 to 13,178.34), MAPE (from 1.80% to 38.20%), and MAE (from 508.49 to 11,654.64), the Helformer achieves significantly better results with an RMSE of 36.23, MAPE of 0.10%, and MAE of 27.86. Similarly, when compared with the advanced hybrid models used by Jin and Li [28], such as VMD-AGRU-RESVMD-LSTM, which recorded an RMSE of 50.651, MAPE of 0.39% and MAE of 42.298, the Helformer demonstrates superior performance with an exceptionally low RMSE of 0.201, MAPE of 0.0014%, and MAE of 0.153. This stark contrast in performance highlights the Helformer's capability to capture complex patterns in time series data with unparalleled precision. Further, the comparison with studies by Seabe et al. [53] and Hansun et al. [23] also underscores Helformer's dominance. These comparisons show that Helformer outperforms both traditional and advanced models used in existing studies, proving its robustness, versatility, and state-of-the-art capability in predicting cryptocurrency prices with far greater accuracy and reliability.

Generalization and transfer learning ability of Helformer

Transfer learning in finance is a methodology that enables the development of high-performance models trained with data from one market and applied to another within the same domain, particularly useful when acquiring sufficient training data is costly or challenging [19]. It allows a model to leverage previously learned knowledge and apply it to a closely related but distinct task, thereby enhancing its overall predictive proficiency. Although transfer learning is still relatively new in cryptocurrency forecasting, its potential to significantly reduce the data and computational resources required for training new models makes it a valuable technique for time series prediction. To implement this approach, the Helformer model was initially trained on the BTC dataset to develop a robust foundational model. Once the optimal model configuration was identified, its generalizability and cross-learning ability were tested by applying the pre-trained model to datasets of the top 15 cryptocurrencies ranked by market capitalization. Without fine-tuning the optimized model parameters, the assessment focused on evaluating its predictive power on different assets without retraining from scratch. The results in Table 8 demonstrate that even without further parameter adjustments, Helformer achieved exceptional predictive accuracy and robustness across multiple cryptocurrencies. This highlights its ability to generalize effectively across different cryptocurrencies, reinforcing its reliability as a versatile forecasting model.

The evaluation metrics for 15 selected cryptocurrencies, using a pre-trained model on BTC, are presented in Table 8. It shows outstanding predictions across various metrics, including RMSE, MAPE, MAE, R^2, EVS, and KGE, reflecting the model's ability to effectively generalize the patterns learned from BTC to other cryptocurrencies. For ETH and BCH, RMSE values are 15.0676 and 10.0356, respectively, indicating some variability in model predictions, yet both show high R^2 and EVS values close to 1, suggesting that the model captures a significant proportion of the variance in these cryptocurrencies. The KGE values for ETH and BCH are 0.9916 and 0.9541, respectively, which are relatively high, demonstrating good agreement between the observed and predicted values. Cryptocurrencies such as SOL and TRX showcase

Table 8 Evaluation metrics of 15 selected stocks using a pre-trained model on BTC

S/N	Cryptocurrency	RMSE	MAPE	MAE	R^2	EVS	KGE
1	ETH	15.0676	0.6039%	14.0754	0.9995	0.9999	0.9916
2	BNB	9.2982	2.4629%	8.5706	0.9957	0.9993	0.9652
3	SOL	2.6935	2.3311%	2.3447	0.9976	0.9994	0.9670
4	XRP	0.0014	0.2644%	0.0014	0.9996	0.9999	0.9962
5	TON	0.0085	0.1771%	0.0076	0.9999	1	0.9974
6	DOGE	0.0001	0.0606%	$0.5919*10^{-4}$	0.9999	0.9999	0.9998
7	ADA	0.0020	0.4564%	0.0018	0.9997	0.9999	0.9935
8	TRX	$0.4755*10^{-10}$	$0.3045*10^{-7}$%	$0.2854*10^{-10}$	1	1	1
9	AVAX	0.4701	1.3067%	0.4270	0.9986	0.9997	0.9813
10	SHIB	$0.4841*10^{-6}$	2.4623%	$0.4338*10^{-6}$	0.9966	0.9993	0.9653
11	DOT	0.1339	1.8510%	0.1258	0.9939	0.9992	0.9738
12	LINK	0.3891	3.0447%	0.3510	0.9936	0.9988	0.9570
13	BCH	10.0356	3.2494%	8.7577	0.9944	0.9986	0.9541
14	LEO	0.1465	3.1268%	0.1424	0.9742	0.9985	0.9558
15	NEAR	0.0461	0.8978%	0.0385	0.9995	0.9998	0.9876

Table 9 Trading results of Helformer model vs B&H strategy

	Trading Strategy	Helformer				B&H			
S/N	Coins	ER (%)	V	MDD	SR	ER (%)	V	MDD	SR
1	ETH	854.88	0.0204	−0.0043	16.46	119.08	0.0272	−0.2456	1.12
2	BNB	493.80	0.0244	−0.0502	7.95	100.95	0.0266	−0.4462	1.01
3	SOL	937.72	0.0371	−0.0358	15.70	612.61	0.0481	−0.1940	2.52
4	XRP	1044.18	0.0331	−0.0007	12.41	27.19	0.0399	−0.3125	0.22
5	TON	668.86	0.0320	−0.0010	19.36	236.66	0.0456	−0.1826	2.45
6	DOGE	1354.79	0.0305	−0.0004	17.51	66.72	0.0418	−0.4040	0.47
7	ADA	1204.52	0.0250	−0.0017	18.93	16.55	0.0356	−0.4839	0.15
8	TRX	656.68	0.0148	0.0000	17.42	86.74	0.0202	−0.1586	1.19
9	AVAX	988.94	0.0352	−0.0061	19.45	219.99	0.0507	−0.3093	1.58
10	SHIB	831.66	0.0555	0.0000	12.26	88.88	0.0666	−0.3144	0.77
11	DOT	692.42	0.0310	−0.0252	15.14	53.96	0.0399	−0.3515	0.72
12	LINK	882.63	0.0345	−0.0350	10.05	108.10	0.0394	−0.4214	0.72
13	BCH	846.55	0.0437	−0.0411	7.62	216.39	0.0474	−0.2667	0.94
14	LEO	167.04	0.0169	−0.0654	5.02	48.53	0.0176	−0.1247	1.12
15	NEAR	1159.39	0.0434	−0.0079	18.87	382.34	0.0614	−0.2062	1.80

impressive model accuracy, with TRX achieving nearly perfect scores across all metrics, highlighting the model's exceptional performance in handling this asset. Overall, the result demonstrates the potential of the Helformer model as a powerful tool for cryptocurrency forecasting, capable of adapting learned behaviours from BTC to a diverse set of other cryptocurrencies.

To further evaluate the trading strategy results of the Helformer model, its performance was compared against the B&H strategy for all the selected cryptocurrencies. Table 9 presents the results, including key performance metrics such as ER, V, MDD, and SR for both strategies across the 15 selected coins. These metrics help assess the trading strategies' profitability and risk management capabilities, revealing that the

Helformer model consistently outperforms the B&H strategy in terms of ER for all 15 cryptocurrencies.

For example, ETH showcases a dramatic improvement in the Helformer model, with an ER of 854.88% and a Sharpe Ratio (SR) of 16.46, which significantly outperforms the B&H strategy's ER of 119.08% and SR of 1.12. This pattern is consistent across other cryptocurrencies, where the Helformer model not only yields higher returns but also demonstrates more efficient risk management. For instance, DOGE presents an ER of 1354.79% and an extremely low MDD of -0.0004, compared to B&H's ER of 66.72% and a higher MDD of -0.4040, illustrating the Helformer's ability to generate substantial returns while minimizing potential losses. The Helformer model also consistently exhibits lower volatility across most cryptocurrencies compared to B&H, indicating a more stable and less risky trading performance. For ADA, the Helformer achieves volatility of 0.0250 compared to 0.0356 for B&H, further highlighting its effectiveness in managing market fluctuations. Additionally, the Helformer achieves remarkably high SR, such as 19.36 for TON and 18.93 for ADA, suggesting a superior risk-adjusted return relative to B&H, which shows considerably lower SR.

This stark contrast in trading performance is further evident in cryptocurrencies like SHIB and AVAX, where the Helformer improves the return and significantly reduces the impact of potential large drawdowns, as seen in the much lower MDD values. For example, AVAX under Helformer experiences an MDD of −0.0061 compared to -0.3093 under B&H, indicating less vulnerability to sudden market downturns. In sum, the Helformer model not only delivers much higher excess returns across all cryptocurrencies but also manages risk more effectively, as evidenced by lower volatility, smaller drawdowns, and higher Sharpe Ratios. These findings confirm the versatility and robustness of the Helformer model in real-world trading scenarios, emphasizing its value as a powerful tool for investors seeking both high returns and controlled risk in the volatile cryptocurrency market.

Conclusion, limitations, and future directions

This work introduces the Helformer model, which represents a significant progression in the field of cryptocurrency price forecasting. The model integrates robust hyperparameter optimization techniques and leverages the strengths of Transformer architectures to tackle the unique challenges presented by highly volatile financial time series like those of cryptocurrencies. By incorporating elements such as Holt-Winters exponential smoothing for time series decomposition and an LSTM component in place of the typical FFN, Helformer adeptly handles non-stationarities and seasonality, features prevalent in cryptocurrency data. The empirical results from extensive tests demonstrate Helformer's superior accuracy and robustness in predicting cryptocurrency prices compared to traditional models. Its capability to generalize across various cryptocurrencies, as evidenced by transfer learning applications, further emphasizes its practical utility and versatility in real-world trading scenarios. The integration of Bayesian optimization with Optuna for hyperparameter tuning also highlights a methodological advancement, improving model reliability and performance. By harnessing cutting-edge deep learning techniques and sophisticated model optimization strategies, the Helformer model addresses the volatile nature of cryptocurrencies, giving room for more stable and predictable investment strategies.

In the future, there are various potential areas for further research and exploration. Firstly, broadening the model's scope to encompass a wider range of financial instruments beyond cryptocurrencies could unlock new markets and opportunities. Investigating the applicability of the Helformer model in other volatile financial markets, such as stock indices, commodities, or Forex markets, would be a valuable extension. Secondly, while the current study focuses on univariate time series forecasting, incorporating multivariate data could significantly enhance the model's predictive accuracy. Future research could integrate technical indicators, sentiment analysis, macroeconomic indicators, and on-chain data to improve decision-making in cryptocurrency and financial market predictions. This would allow the model to capture external influences that impact price movements and market behavior. Third, exploring deeper integrations with reinforcement learning could refine the model's trading strategy component. This approach could evolve Helformer from merely predicting prices to actively suggesting and managing dynamic trading strategies, potentially increasing profitability and minimizing risks in real-time trading environments.

Additionally, while the present study focuses on next-day price forecasting, future studies should investigate multi-step or multi-horizon forecasting, where predictions extend beyond a single time step. Since longer prediction windows often introduce more uncertainty and higher error rates, evaluating Helformer's performance in long-term forecasting scenarios would provide further insights into its generalization capability and limitations. By pursuing these future directions, the Helformer model can continue to lead in technological innovation while promoting a responsible, adaptable, and equitable financial technology landscape.

Acknowledgements
The authors are grateful to the Hong Kong Polytechnic University for financial and technical support.

Author contributions
T.O. Kehinde: Conceptualization, Methodology, Writing – original draft, Software. Oluyinka J. Adedokun: Writing – review and editing, Investigation, Validation. Akpan Joseph: Formal Analysis, Resources. Kareem Morenikeji Kabirat: Software, Visualization, Investigation. Hammed Adebayo Akano: Validation, Data Curation. Oludolapo A. Olanrewaju: Supervision, Project Administration, Funding Acquisition. All authors reviewed the manuscript.

Funding
The authors are grateful to the Hong Kong Polytechnic University for financial and technical support.

Availability of data and materials
Data is available upon reasonable request.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 December 2024 Accepted: 25 March 2025
Published online: 03 April 2025

References
1. Abu Bakar N, Rosbi S. Autoregressive integrated moving average (ARIMA) model for forecasting cryptocurrency exchange rate in high volatility environment: a new insight of bitcoin transaction. Int J Adv Eng Res Sci. 2017;4(11):130–7.

2. Akyildirim E, Goncu A, Sensoy A. Prediction of cryptocurrency returns using machine learning. Ann Oper Res. 2021;297:3–36.
3. Alonso-Monsalve S, Suárez-Cetrulo AL, Cervantes A, Quintana D. Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators. Expert Syst Appl. 2020;149: 113250.
4. Amadeo, A. J., Siento, J. G., Eikwine, T. A., & Parmonangan, I. H. Temporal Fusion Transformer for Multi Horizon Bitcoin Price Forecasting. 2023 IEEE 9th Information Technology International Seminar (ITIS), 2023
5. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint* arXiv:2004.05150.
6. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13(2):2.
7. Bouteska A, Abedin MZ, Hajek P, Yuan K. Cryptocurrency price forecasting–a comparative analysis of ensemble learning and deep learning methods. Int Rev Financ Anal. 2024;92: 103055.
8. Catania L, Grassi S, Ravazzolo F. Forecasting cryptocurrencies under model and parameter instability. Int J Forecast. 2019;35(2):485–501.
9. Cavalli S, Amoretti M. CNN-based multivariate data analysis for bitcoin trend prediction. Appl Soft Comput. 2021;101: 107065.
10. Chang T-J, Lee T-S, Yang C-T, Lu C-J. A ternary-frequency cryptocurrency price prediction scheme by ensemble of clustering and reconstructing intrinsic mode functions based on CEEMDAN. Expert Syst Appl. 2023;233: 121008.
11. Chowdhury R, Rahman MA, Rahman MS, Mahdy M. An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning. Physica A. 2020;551: 124569.
12. Conrad C, Custovic A, Ghysels E. Long-and short-term cryptocurrency volatility components: a GARCH-MIDAS analysis. J Risk Financ Manag. 2018;11(2):23.
13. Da Silva, R. G., Ribeiro, M. H. D. M., Fraccanabbia, N., Mariani, V. C., & dos Santos Coelho, L. Multi-step ahead bitcoin price forecasting based on VMD and ensemble learning methods. 2020 International Joint Conference on Neural Networks (IJCNN). 2020
14. Du X, Tang Z, Wu J, Chen K, Cai Y. A new hybrid cryptocurrency returns forecasting method based on multiscale decomposition and an optimized extreme learning machine using the sparrow search algorithm. Ieee Access. 2022;10:60397–411.
15. Dutta A, Kumar S, Basu M. A gated recurrent unit approach to bitcoin price prediction. J Risk Financ Manag. 2020;13(2):23.
16. Fallah MF, Pourmansouri R, Ahmadpour B. Presenting a new deep learning-based method with the incorporation of error effects to predict certain cryptocurrencies. Int Rev Financ Anal. 2024;95: 103466.
17. Ghosh I, Jana RK, Sharma DK. A novel granular decomposition based predictive modeling framework for cryptocurrencies' prices forecasting. China Financ Rev Int. 2024. https://doi.org/10.1108/CFRI-03-2023-0072.
18. Girsang AS. Hybrid LSTM and GRU for cryptocurrency price forecasting based on social network sentiment analysis using FinBERT. Ieee Access. 2023;11:120530–40.
19. Golnari A, Komeili MH, Azizi Z. Probabilistic deep learning and transfer learning for robust cryptocurrency price prediction. Expert Syst Appl. 2024. https://doi.org/10.1016/j.eswa.2024.124404.
20. Goodell JW, Jabeur SB, Saâdaoui F, Nasir MA. Explainable artificial intelligence modeling to forecast bitcoin prices. Int Rev Financ Anal. 2023;88: 102702.
21. Hamayel MJ, Owda AY. A novel cryptocurrency price prediction model using GRU, LSTM and bi-LSTM machine learning algorithms. Ai. 2021;2(4):477–96.
22. Hanifi S, Cammarono A, Zare-Behtash H. Advanced hyperparameter optimization of deep learning models for wind power prediction. Renew Energy. 2024;221: 119700.
23. Hansun S, Wicaksana A, Khaliq AQ. Multivariate cryptocurrency prediction: comparative analysis of three recurrent neural networks approaches. J Big Data. 2022;9(1):50.
24. Haryono AT, Sarno R, Sungkono KR. Transformer-gated recurrent unit method for predicting stock price based on news sentiments and technical indicators. Ieee Access. 2023. https://doi.org/10.1109/ACCESS.2023.3298445.
25. Ho K-H, Hou Y, Georgiades M, Fong KC. Exploring key properties and predicting price movements of cryptocurrency market using social network analysis. Ieee Access. 2024. https://doi.org/10.1109/ACCESS.2024.3397723.
26. Ibrahim A, Kashef R, Corrigan L. Predicting market movement direction for bitcoin: A comparison of time series modeling methods. Comput Electr Eng. 2021;89: 106905.
27. Jay P, Kalariya V, Parmar P, Tanwar S, Kumar N, Alazab M. Stochastic neural networks for cryptocurrency price prediction. Ieee Access. 2020;8:82804–18.
28. Jin C, Li Y. Cryptocurrency price prediction using frequency decomposition and deep learning. Fractal Fract. 2023;7(10):708.
29. Kehinde T, Chan FT, Chung S. Scientometric review and analysis of recent approaches to stock market forecasting: two decades survey. Expert Syst Appl. 2023;213: 119299.
30. Kehinde T, Chung S, Chan FT. Benchmarking TPU and GPU for Stock Price Forecasting Using LSTM Model Development. In: Science and information conference. Cham: Springer; 2023.
31. Koo E, Kim G. Centralized decomposition approach in LSTM for Bitcoin price prediction. Expert Syst Appl. 2024;237: 121401.
32. Kristjanpoller W, Minutolo MC. A hybrid volatility forecasting framework integrating GARCH, artificial neural network, technical analysis and principal components analysis. Expert Syst Appl. 2018;109:1–11.
33. Kumarappan J, Rajasekar E, Vairavasundaram S, Kotecha K, Kulkarni A. Siamese graph convolutional split-attention network with NLP based social sentimental data for enhanced stock price predictions. J Big Data. 2024;11(1):154.
34. Li J, Zhang Y, Yang X, Chen L. Online portfolio management via deep reinforcement learning with high-frequency data. Inf Process Manage. 2023;60(3): 103247.
35. Li Y, Jiang S, Li X, Wang S. Hybrid data decomposition-based deep learning for bitcoin prediction and algorithm trading. Financ Innov. 2022;8(1):31.
36. Liu M, Li G, Li J, Zhu X, Yao Y. Forecasting the price of Bitcoin using deep learning. Financ Res Lett. 2021;40: 101755.

37. Livieris IE, Kiriakidou N, Stavroyiannis S, Pintelas P. An advanced CNN-LSTM model for cryptocurrency forecasting. Electronics. 2021;10(3):287.
38. Lu, Y., Zhang, H., & Guo, Q. (2023). Stock and market index prediction using Informer network. arXiv preprint arXiv:2305.14382.
39. Misra, D. (2019). Mish: A self regularized non-monotonic activation function. arXiv preprint arXiv:1908.08681.
40. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Satoshi Nakamoto.
41. Nakano M, Takahashi A, Takahashi S. Bitcoin technical trading with artificial neural network. Physica A. 2018;510:587–609.
42. Nasirtafreshi I. Forecasting cryptocurrency prices using recurrent neural network and long short-term memory. Data Knowl Eng. 2022;139: 102009.
43. Otabek S, Choi J. From prediction to profit: a comprehensive review of cryptocurrency trading strategies and price forecasting techniques. Ieee Access. 2024. https://doi.org/10.1109/ACCESS.2024.3417449.
44. Oyedele AA, Ajayi AO, Oyedele LO, Bello SA, Jimoh KO. Performance evaluation of deep learning and boosted trees for cryptocurrency closing price prediction. Expert Syst Appl. 2023;213: 119233.
45. Oyewola DO, Dada EG, Ndunagu JN. A novel hybrid walk-forward ensemble optimization for time series cryptocurrency prediction. Heliyon. 2022. https://doi.org/10.1016/j.heliyon.2022.e11862.
46. Patel MM, Tanwar S, Gupta R, Kumar N. A deep learning-based cryptocurrency price prediction scheme for financial institutions. J Inf Security Appl. 2020;55: 102583.
47. Peng P, Chen Y, Lin W, Wang JZ. Attention-based CNN–LSTM for high-frequency multiple cryptocurrency trend prediction. Expert Syst Appl. 2024;237: 121520.
48. Poongodi M, Nguyen TN, Hamdi M, Cengiz K. Global cryptocurrency trend prediction using social media. Inf Process Manag. 2021;58(6): 102708. https://doi.org/10.1016/j.ipm.2021.102708.
49. Quan SJ. Comparing hyperparameter tuning methods in machine learning based urban building energy modeling: a study in Chicago. Energy Build. 2024. https://doi.org/10.1016/j.enbuild.2024.114353.
50. Rathore RK, Mishra D, Mehra PS, Pal O, Hashim AS, Shapi'i A, Ciano T, Shutaywi M. Real-world model for bitcoin price prediction. Inf Process Manag. 2022;59(4):102968. https://doi.org/10.1016/j.ipm.2022.102968.
51. Saheed YK, Kehinde TO, Ayobami Raji M, Baba UA. Feature selection in intrusion detection systems: a new hybrid fusion of Bat algorithm and Residue Number System. J Inf Telecommun. 2024;8(2):189–207.
52. Sbrana A, Lima de Castro PA. N-BEATS perceiver: a novel approach for robust cryptocurrency portfolio forecasting. Comput Econ. 2023;2:1–35.
53. Seabe PL, Moutsinga CRB, Pindza E. Forecasting cryptocurrency prices using LSTM, GRU, and bi-directional LSTM: a deep learning approach. Fractal Fract. 2023;7(2):203.
54. Sebastião H, Godinho P. Forecasting and trading cryptocurrencies with machine learning under changing market conditions. Financ Innov. 2021;7:1–30.
55. Smyl S. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. Int J Forecast. 2020;36(1):75–85.
56. Sun X, Liu M, Sima Z. A novel cryptocurrency price trend forecasting model based on LightGBM. Financ Res Lett. 2020;32: 101084.
57. Tanwar, A., & Kumar, V. (2022). Prediction of cryptocurrency prices using transformers and long short term neural networks. 2022 International Conference on Intelligent Controller and Computing for Smart Power (ICICCSP),
58. Touzani Y, Douzi K. An LSTM and GRU based trading strategy adapted to the Moroccan market. J Big Data. 2021;8(1):126.
59. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:1.
60. Walther T, Klein T, Bouri E. Exogenous drivers of Bitcoin and Cryptocurrency volatility–a mixed data sampling approach to forecasting. J Int Finan Markets Inst Money. 2019;63: 101133.
61. Wu H, Xu J, Wang J, Long M. Autoformer: decomposition transformers with auto-correlation for long-term series forecasting. Adv Neural Inf Process Syst. 2021;34:22419–30.
62. Zhang Z, Dai H-N, Zhou J, Mondal SK, García MM, Wang H. Forecasting cryptocurrency price using convolutional neural networks with weighted and attentive memory channels. Expert Syst Appl. 2021;183: 115378.
63. Zhong C, Du W, Xu W, Huang Q, Zhao Y, Wang M. LSTM-ReGAT: a network-centric approach for cryptocurrency price trend prediction. Decis Support Syst. 2023;169: 113955.
64. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, Zhang W. Informer: Beyond efficient transformer for long sequence time-series forecasting. Proc AAAI Conf Artif Intell. 2021. https://doi.org/10.1609/aaai.v35i12.17325.
65. Zhou T, Ma Z, Wen Q, Wang X, Sun L, Jin R. Fedformer: frequency enhanced decomposed transformer for long-term series forecasting. Int Conf Mach Learn. 2022;162:27268.
66. Zhou Z, Song Z, Xiao H, Ren T. Multi-source data driven cryptocurrency price movement prediction and portfolio optimization. Expert Syst Appl. 2023;219: 119600.
67. Zoumpekas T, Houstis E, Vavalis M. ETH analysis and predictions utilizing deep learning. Expert Syst Appl. 2020;162: 113866.

Publisher's Note