FORECASTING STOCK PRICES IN VIETNAM USING MACHINE LEARNING AND DEEP LEARNING MODELS

TRAN TRUC QUYNH[1], LUONG THI THUY DIEM[2], AND NGUYEN HUU THANH[3]

[1]Faculty of Information Systems, University of Information Technology, (e-mail: 21522539@gm.uit.edu.vn)
[2]Faculty of Information Systems, University of Information Technology, (e-mail: 21521953@gm.uit.edu.vn)
[3]Faculty of Information Systems, University of Information Technology, (e-mail: 21522599@gm.uit.edu.vn)

ABSTRACT The stock market has assumed a significant position within investment portfolios, as evidenced by the recent surge in investor participation within Vietnam's market. However, navigating the inherent volatility of stock prices requires experience and robust analytical tools. This paper seeks to empower investors by employing various forecasting models such as Linear Regression (LR), Auto Regressive Integrated Moving Average (ARIMA), ecurrent neural network (RNN), Gated Recurrent Units (GRU),Long short term memory (LSTM), Fast Fourier Transform (FFT),XGBOOST, Fully Convolutional Networks (FCNs) to analyze price trends within three prominent bank stocks: ACB, BIDV, VCB traded on Hochiminh Stock Exchange (HOSE) over a five-year period (2019-2024). By evaluating the effectiveness of these models using RMSE, MAPE, and MSLE techniques. This paper aims to equip investors for informed stock selection within their portfolios, ultimately contributing to more accurate decision-making.

INDEX TERMS Time Series Analysis, Machine Learning, Financial Forecasting, Vietnamese Stock Market, GRU, RNN, ARIMA, LSTM, FFT, XGBoost, FCN, Linear Regression.

I. INTRODUCTION

The Vietnamese stock market has witnessed a record-high level of participation in recent years, reflecting the growing attention and interest it has garnered. This trend is a positive indicator of the market's potential and the overall recovery of the economy in the post-2023 period. However, the stock market also inherently carries short-term risks that investors must navigate. Among the various sectors, the banking industry has traditionally occupied a high proportion and serves as a crucial pillar of the economy.

The purpose of this paper is to support investors in making accurate decisions regarding three bank stocks (ACB, BID, VCB) listed on the Hochiminh Stock Exchange (HOSE). To achieve this goal, the study will apply various models, including Linear Linear Regression (LR), Auto Regressive Integrated Moving Average (ARIMA), ecurrent neural network (RNN), Gated Recurrent Units (GRU),Long short term memory (LSTM), Fast Fourier Transform (FFT),XGBOOST, Fully Convolutional Networks (FCNs) to the stock price data from 2019 to 2024 to forecast the trends.Techniques such as RMSE, MAPE, and MSLE will be utilized to evaluate the effectiveness of these models.

II. RELATED WORKS

Over the past several years, a significant volume of research has concentrated on forecasting stock prices by leveraging various machine learning and statistical techniques. Both CAKRA (2015) and URAS (2020) used linear regression to forecast stock prices and Bitcoin closing prices. CAKRA applied linear regression based on sentiment analysis to process Twitter data, improving the accuracy of stock price predictions [1]. Meanwhile, URAS combined linear regression and neural network models to forecast Bitcoin closing prices, showing that linear regression had faster execution speed and could accurately predict Bitcoin price fluctuations [2].

Manish Dadhich studied and applied the ARIMA model for short-term forecasting of BSE and NSE stock prices. After carrying out his research, Manish Dadhich demonstrated the strength of the ARIMA model in predicting daily closing prices of time series data [3]. In Anusha Garlapati's research, she also used ARIMA to forecast stock prices and concluded that ARIMA is a good model for predicting stock prices [4].

Yongqiong Zhu [5] applied an RNN model to predict the stock prices of Apple. The training dataset spanned 10 years with 65% allocated for training and the remaining 3%

for testing. With 50 epochs, Adam optimization, and Mean Squared Error (MSE) as the loss function, the model achieved highly favorable outcomes. It attained a predictive accuracy exceeding 95%, with a reported loss value of 0.1%. In 2015, Xiao Ding [6] also proposed a deep learning method for event-driven stock market prediction.

Shejul et al. [7] compared the performance of the Gated Recurrent Unit (GRU) and Bidirectional Long Short-Term Memory (BiLSTM) models in predicting stock prices. Experimental results indicate that both models accurately forecast future stock prices, with the BiLSTM model outperforming the GRU model. Despite this, the GRU model demonstrates nearly double the speed of the BiLSTM model due to its simpler architecture. Overall, both models offer precise predictions and can effectively anticipate future stock market trends.

C. Fjellström (2022) [8] explored an LSTM ensemble for stock price prediction in the paper "Long Short-Term Memory Neural Network for Financial Time Series." The work demonstrates superior performance over traditional portfolios, offering insights into LSTM's potential for financial forecasting and strategies for enhancing model accuracy and reducing market risk. S. Mehtab, J. Sen (2020) [9] proposed a deep learning approach for stock price prediction by employing Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. Their work contributes to the field by exploring the effectiveness of this combined deep learning architecture in forecasting the NIFTY 50 index prices. They further emphasize the importance of data preprocessing and model evaluation, providing a methodological blueprint for financial market analysis using deep learning techniques.

Chen et al. [10] used an Fast Fourier Transform algorithm to deal with historical training data for forecasting stock prices, achieving a more highly accurate. Experimental analysis is performed on datasets covering a seven-year period of both the Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) and the Dow-Jones Industrial Average (DJIA).

A. Qingwen Jin et al. [11] established predictive models using the Best Track TC dataset and the XGBOOST algorithm to anticipate Tropical Cyclone (TC) intensity in the Western North Pacific (WNP). Across six scenarios, the model achieved high accuracy with MAE < 4.50 m/s, CC > 0.89, and NRMSE < 10.00%. The XGBOOST model exhibited superior performance compared to traditional Back-Propagation Neural Network (BPNN) models for the same predictors and independent prediction samples. A. O. A. A. A. Tianqi Chen and Carlos Guestrin (2016) [12] introduced XGBoost, a scalable tree boosting system. The paper describes the architecture of XGBoost and its core algorithms, presenting improvements to enhance performance and accuracy across a variety of machine learning tasks.

Fully Convolutional Networks (FCNs) have been studied and applied in various fields related to image data processing and semantic segmentation. According to Shima Nabiee's research on stock trend prediction, FCNs have also been proven to be a powerful and flexible tool, enabling the analysis and prediction of stock price trends based on raw data [13].

III. MATERIALS

A. DATASET

The analysis will focus on the historical stock prices of three banks in Vietnam: the Asia Commercial Joint Stock Bank (ACB), the Bank for Investment and Development of Vietnam (BIDV), and the Joint Stock Commercial Bank for Foreign Trade of Vietnam (VCB). The data spans from January 3, 2019, to January 3, 2024, and includes information such as date, price, opening price, highest and lowest prices, volume, and price change. However, the primary aim is to forecast closing prices, so only the "Price" column (VND) will be used for analysis.

B. DESCRIPTIVE STATISTICS

TABLE 1. ACB, BIDV, VCB's Descriptive Statistics

	ACB	BID	VCB
Count	1247	1252	1252
Mean	19.712	35.993	74.810
Standard Deviation	6.205	6.574	12.664
Min	8.763	23.420	43.925
25%	11.963	31.226	65.274
50%	22.000	34.823	75.871
75%	24.980	41.600	84.525
Max	30.360	53.900	106.500
Variance	49.100	28.667	106.500
Skewness	-0.345	0.299	-0.132
Kurtosis	-1.457	-0.821	-0.530

1) ACB stock price visualization

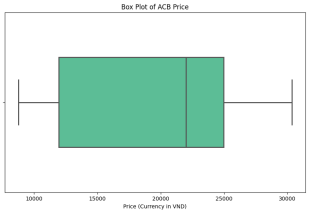

FIGURE 1. ACB stock price's boxplot

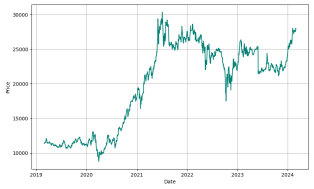

FIGURE 2. ACB stock price's time

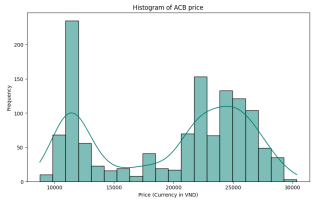

FIGURE 3. ACB stock price's histogram

ACB's price distribution appears left-skewed, as evidenced by the pronounced peak in the histogram at the lower end of the price range. Interestingly, ACB's box plot appears relatively compact, with the median closer to the higher end of the price range

2) BID stock price visualization

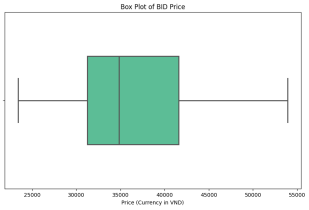

FIGURE 4. BID stock price's boxplot

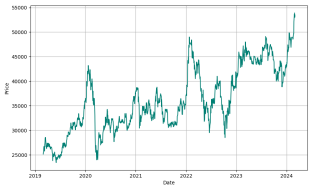

FIGURE 5. BID stock price's time

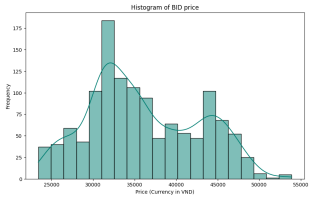

FIGURE 6. BID stock price's histogram

The value of BID shares mainly fluctuates between 33,000 VND and 43,000 VND, and the highest current value that BID shares have reached is 54,000 VND. Additionally, the current trend regarding the value of BID shares is on the rise.

3) VCB stock price visualization

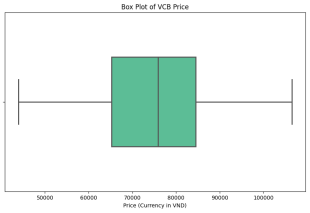

FIGURE 7. VCB stock price's boxplot

FIGURE 8. VCB stock price's time

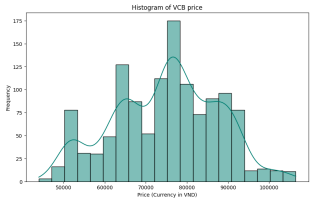

FIGURE 9. VCB stock price's histogram

Based on VCB's Boxplot, it can be seen that most of the data is concentrated in Q2 to Q3 with length 8654 approximately 75871 to 84525. The highest concentration on the chart at the price of about 76,000 VND with a frequency of nearly 175 times may be a sign of investors' special interest in VCB stock price. Besides, the price also fluctuates frequently around 70,000 The data chart of price fluctuations over the years shows that stock prices tend to gradually increase, although prices also decrease to the lowest (43925) in early 2020 and the highest (106500) in mid-2023.

IV. METHODOLOGY

A. ARIMA

Auto Regressive Integrated Moving Average (ARIMA) is a model that describes time series based on observed values, which can be used to forecast future values. Applying ARIMA models to any time series showing patterns with no random white noise and non-seasonality. The model was introduced by Box and Jenkins in 1970. To generate short-term forecasts, ARIMA models have shown efficient capabilities, outperforming complex structural models. The future value of a variable in the ARIMA model is a combination of linearity to the past values and errors, expressed as follows [15]:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \ldots - \theta_q \varepsilon_{t-q}$$

Where:

- Y_t is the actual value at time t.
- ε_t is the random error at time t.
- ϕ_i and θ_j are the coefficients.
- p and q are integers, often referred to as autoregressive and moving average parameters, respectively.

B. FAST FOURIER TRANSFORM

The Fast Fourier Transform (commonly abbreviated as FFT) is a fast algorithm for computing the discrete Fourier transform (DFT) of a sequence [15] by using the factor N/2 log N where N is the number of points, therefore, the FFT is a way to convert the time series from time domain to frequency domain [16] with equation [17]:

$$F[n] = \sum_{k=0}^{N-1} f[k] \cdot e^{-i\frac{2\pi}{N}kn}$$

where:

- $F[n]$ is the Discrete Fourier Transform of the sequence.
- $f[k]$ is the k-th element of the input sequence.
- N is the total number of elements in the input sequence.
- n is the index of the frequency component in the output sequence $F[n]$.
- k is the index of the element in the input sequence $f[k]$.

C. LINEAR REGRESSION

Linear regression is a statistical technique used to model the relationship between a dependent variable, *Y*, and one or more independent variables, *X*. The goal is to find the best-fitting straight line (or hyperplane in higher dimensions) that describes the relationship between the variables. When there are multiple independent variables, the linear regression is

called multivariable linear regression, with equation has the form [18]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Where:

- Y is the dependent variable.
- $X_1, X_2, \ldots, X_k$ are the independent variables.
- β_0 is the intercept term.
- $\beta_1, ..., \beta_k$ are the regression coefficients for the independent variables.
- ε is the error term.

D. XGBOOST

XGBoost is a highly efficient and scalable implementation of gradient boosting, a powerful machine learning technique. XGBoost operates by constructing a series of decision trees in an additive manner. Each tree is built sequentially, with each one correcting the errors of the previous trees. The model's objective is to minimize the loss function. Key advantages of XGBoost include its ability to handle large datasets through parallelized computing, effective handling of missing data values, and flexible parameter tuning. These features, along with its state-of-the-art algorithm for supervised learning problems, known for its high accuracy, which rivals deep learning models [19] .Unlike deep learning that typically requires numerical raw data, XGBoost can handle tabular datasets of any size and type, including categorical data often found in business models.

Explanation of the mathematics behind XGBoost:

Ensemble Model:XGBoost combines predictions from multiple decision trees (f_k) to create a final prediction (F(x)) [19]:

$$y_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F},$$

Where: $F = f(x) = w_q(x) \mid q : \mathbb{R}^m \to T, w \in \mathbb{R}^T$

Objective Function: It minimizes a loss function (L) that measures prediction error, with a regularization term (Ω) to prevent overfitting: [20]

$$\text{obj}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where, first term is the loss function and the second is the regularization parameter

E. GRU

The Gated Recurrent Unit, just like the LSTM, is a Recurrent Neural Network. It, however, has a less complicated structure compared to LSTM. It lacks an output gate but has an update z and a reset gate r. These gates are vectors which decide what information should be passed to the output. The Reset gate defines how to combine the new input with the previous memory. The definition of how much of the last memory to keep is done by the Update [12]. The GRU has the following equations: [21]

Update gate:

$$z_t = \sigma(W_z h_{t-1} + U_z x_t) \tag{1}$$

Reset gate:

$$r_t = \sigma(W_r h_{t-1} + U_r x_t) \tag{2}$$

Cell state:

$$c_t = \tanh(W_c(h_{t-1} * r_t) + U_c x_t) \tag{3}$$

New state:

$$h_t = (z_t * c_t) + ((1 - z_t) * h_{t-1}) \tag{4}$$

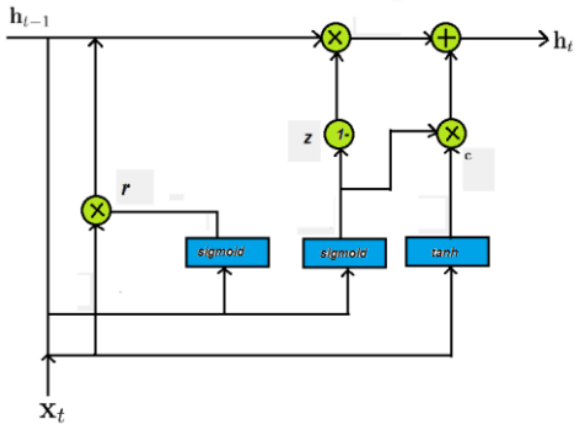

FIGURE 10. Fully Convolutional Neural Network architecture

F. FULLY CONVOLUTIONAL NEURAL NETWORK(FCN)

Fully Convolutional Neural Networks (FCNs) were introduced by Wang et al. (2017b) for univariate time series classification, validated on 44 UCR/UEA datasets. FCNs use convolutional layers without local pooling, maintaining the time series length throughout. A key feature is replacing the final fully connected layer with a Global Average Pooling (GAP) layer, reducing parameters and enabling Class Activation Mapping (CAM) to identify crucial input segments. FCNs lack pooling and regularization, benefiting from parameter invariance across varying time series lengths due to GAP, supporting transfer learning for model adaptation from one dataset to another. [22]

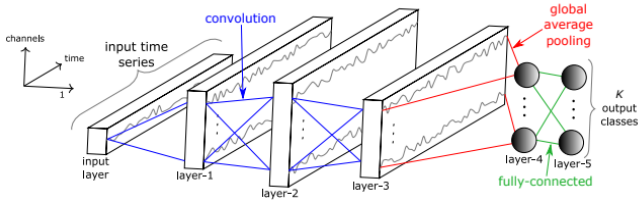

FIGURE 11. Fully Convolutional Neural Network architecture

G. RECURRENT NEURAL NETWORK (RNN)

A recurrent neural network (RNN) is a type of artificial neural network which uses sequential data or time series data [23]. A recurrent neural network (RNN) is an extension of a conventional feedforward neural network, which is able to handle a variable-length sequence input. The RNN handles the variable-length sequence by having a recurrent hidden state whose activation at each time is dependent on that of the previous time [24]. For each timestep t, the activation $a^{<t>}$ and the output $y^{<t>}$ are expressed as follows:

$$a^{<t>} = g_1\left(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a\right)$$

and

$$y^{<t>} = g_2\left(W_{ya}a^{<t>} + b_y\right)$$

where $W_{ax}, W_{aa}, W_{ya}, b_a, b_y$ are coefficients that are shared temporally and g_1, g_2 are activation functions [25].

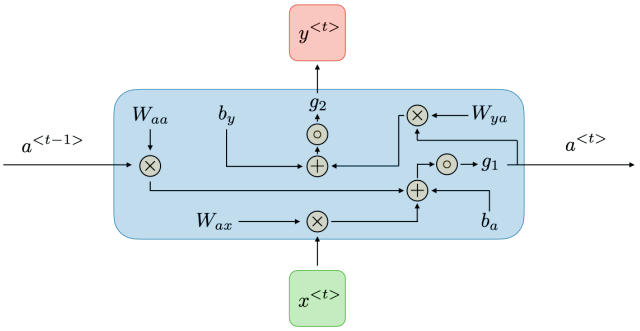

FIGURE 12. Architecture of a Traditional RNN

H. LONG SHORT TERM MEMORY (LSTM)

The Long Short-Term Memory (LSTM) model is a recurrent neural network designed to handle sequential data like time series. Unlike standard RNNs, LSTMs employ memory cells and gating mechanisms to selectively retain and discard information over long sequences. This allows LSTMs to effectively capture long-range dependencies within the sequential data.

At the core of an LSTM are the memory cells regulated by gates - input gate (controls inflow), forget gate (controls clearing), and output gate (controls outflow). This gating architecture enables LSTMs to learn and leverage long-term patterns and relationships present in sequences. By addressing long-range temporal dependencies, LSTMs excel at sequence prediction tasks, making them well-suited for applications such as time series forecasting.

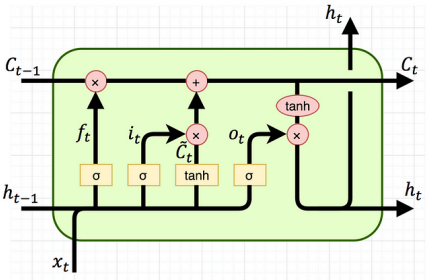

FIGURE 13. LSTM model

Input gate (i_t):

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

Forget gate (f_t):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Output gate (o_t):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

Memory cell (C_t):

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Hidden state (h_t):

$$h_t = o_t * \tanh(C_t)$$

- i_t, f_t, o_t are the values of the gates at time t.
- W_i, W_f, W_o, W_C are weight matrices.
- b_i, b_f, b_o, b_c are bias vectors.
- h_{t-1} is the hidden state from the previous layer.
- x_t is the input at time t.
- C_t is the memory cell state at time t.
- h_t is the hidden state at time t.

REFERENCES

[1] Yahya Eru Cakra and Bayu Distiawan Trisedya. "Stock price prediction using linear regression based on sentiment analysis." In: 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2015. pp. 147-154.

[2] Nicola Uras, et al. "Forecasting Bitcoin closing price series using linear regression and neural networks models." PeerJ Computer Science, vol. 6, 2020, article e279.

[3] Manish Dadhich, et al. “Predictive models for stock market index using stochastic time series ARIMA modeling in emerging economy.” In: Advances in Mechanical Engineering: Select Proceedings of CAMSE 2020. Springer Singapore, 2021. pp. 281-290.

[4] Anusha Garlapati, et al. “Stock price prediction using Facebook Prophet and ARIMA models.” In: 2021 6th International Conference for Convergence in Technology (I2CT). IEEE, 2021. pp. 1-7.

[5] Zhu Y. Stock price prediction using the RNN model. InJournal of Physics: Conference Series 2020 Oct 1 (Vol. 1650, No. 3, p. 032103). IOP Publishing.

[6] Ding X, Zhang Y, Liu T, Duan J. Deep learning for event-driven stock prediction. InTwenty-fourth international joint conference on artificial intelligence 2015 Jun 25.

[7] Shejul, A.A., Chaudhari, A., Dixit, B.A., Lavanya, B.M. (2023). Stock Price Prediction Using GRU, SimpleRNN and LSTM. In: Kulkarni, A.J., Mirjalili, S., Udgata, S.K. (eds) Intelligent Systems and Applications. Lecture Notes in Electrical Engineering, vol 959. Springer, Singapore

[8] Carmina Fjellström. "Long Short-Term Memory Neural Network for Financial Time Series." arXiv:2201.08218v1 [q-fin.ST], 20 Jan 2022.

[9] Sidra Mehtab and Jaydip Sen. "Stock Price Prediction Using CNN and LSTM-Based Deep Learning Mode." Department of Data Science, Praxis Business School, Kolkata, INDIA. Accepted version of paper presented at 2020 International Conference on Decision Aid Sciences and Applications (DASA'20), Bahrain, November 8 – 9, 2020.

[10] Chen MY, Chen BT. Online fuzzy time series analysis based on entropy discretization and a Fast Fourier Transform. Applied Soft Computing. 2014 Jan 1;14:156-66.

[11] Qingwen Jin, Xiangtao Fan, Jian Liu, Zhuxin Xue, and Hongdeng Jian. "Using eXtreme Gradient BOOSTing to Predict Changes in Tropical Cyclone Intensity over the Western North Pacific." Atmosphere, vol. 10, no. 6, 2019, pp. 341–.

[12] Tianqi Chen (University of Washington) and Carlos Guestrin (University of Washington). "XGBoost: A Scalable Tree Boosting System."

[13] Shima Nabiee and Nader Bagherzadeh. "Stock Trend Prediction: A Semantic Segmentation Approach." arXiv preprint arXiv:2303.09323, 2023.

[14] Gillian Smith, " The Fast Fourier Transform and its Applications ", https://www.maths.ed.ac.uk/~ateckent/vacation_reports/summer_project_gillian_smith.pdf, Accessed August 2019.

[15] Gillian Smith, " The Fast Fourier Transform and its Applications ",https://www.maths.ed.ac.uk/~ateckent/vacation_reports/summer_project_gillian_smith.pdf, Accessed August 2019.

[16] Musbah, H., El-Hawary, M., & Aly, H. (2019). Identifying Seasonality in Time Series by Applying Fast Fourier Transform. 2019 IEEE Electrical Power and Energy Conference (EPEC).

[17] S.J. Roberts. "Lecture 7 - The Discrete Fourier Transform". Available at: https://www.robots.ox.ac.uk/~sjrob/Teaching/SP/l7.pdf. Accessed: 2000.

[18] Evans, J. R., *Business Analytics: Methods, Models, and Decisions*. Hoboken, NJ: Wiley, 2013, p. 276.

[19] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," arXiv:1603.02754v3 [cs.LG], June 10, 2016.

[20] P. Pawangfg, "XGBoost," GeeksforGeeks, Available: https://www.geeksforgeeks.org/xgboost/. [Accessed: 06-Feb-2023].

[21] Yamak, P. T., Yujian, L., & Gadosey, P. K. (2019, December). A comparison between arima, lstm, and gru for time series forecasting. In Proceedings of the 2019 2nd international conference on algorithms, computing and artificial intelligence (pp. 49-55).

[22] Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: a review. Data mining and knowledge discovery, 33(4), 917-963.

[23] IBM Technology company, "What are recurrent neural networks?", Available at: https://www.ibm.com/topics/recurrent-neural-networks

[24] Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling." arXiv preprint arXiv:1412.3555 (2014).

[25] A. Amidi and S. Amidi, "Recurrent Neural Networks Cheatsheet," Stanford University, [Online]. Available: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks. [Accessed: 29-May-2024].