DOI:10.15625/vap.2025.0046

DETECTION OF STOCK PRICE ANOMALIES IN VIETNAM MARKET USING THE LSTM AUTO-ENCODER AND ENSEMBLE STACKING METHODS

Nguyen Trung Dung[1,3*], Thai Minh Hanh[2,3*], Dang Van Thang[1], Pham Thu Trang[1]

[1]*Faculty of Mathematics and Informatics, Hanoi University of Science and Technology, Hanoi, Vietnam*
[2]*School of Economics and Management, Hanoi University of Science and Technology, Hanoi, Vietnam*
[3]*BKFintech, Hanoi University of Science and Technology, Hanoi, Vietnam*
**Corresponding author: dung.nguyentrung2@hust.edu.vn*

ABSTRACT

Stock price manipulation is one of the biggest issues in the stock market. In fact, there have been many analytical methods including methods based on machine learning algorithms or deep learning models for this problem. Anomaly detection in the stock market is also an important and challenging problem, especially for Vietnam Stock Exchange (VNINDEX). This problem can be solved by using data analysis methods and artificial intelligence to find patterns and trends in trading activities. The main contribution of this paper is to implement the univariate - Long Short-Term Memory (LSTM) method for stock price prediction and utilize the unsupervised learning model LSTM-based autoencoder, and the supervised ensemble model for anomaly detection and classification. The empirical results showed that our models could detect the anomalies/manipulation in both synthetic and real short-time stock price data in Vietnam.

Research purpose: Anomaly detection using the unsupervised learning model LSTM-based autoencoder, and the supervised ensemble model.

Research motivation: Analyzing stock data to help investors understand the market situation and make effective investment decisions.

Research design, approach, and method: Collecting historical short stock market data from Vietnam and implementing the univariate-LSTM method for stock price prediction and utilize the unsupervised learning model LSTM-based autoencoder, and the supervised ensemble model for anomaly detection and classification.
Main findings: Build a LSTM-based autoencoder model, and an ensemble model for anomaly detection in short-time data from Vietnamese stock market.
Practical/Managerial implications: Matlab and Python codes were created to and implementing the univariate-LSTM method for stock price prediction and utilize the unsupervised learning model LSTM-based autoencoder.

Keywords: *LSTM-based autoencoder, ensemble model, anomaly detection, stock market of Vietnam.*

1. INTRODUCTION

Analyzing stock data is crucial in the field of finance and investment to help investors understand the market situation and make effective investment decisions. It’s also important from the

administrator's perspective to detect and effectively handle the illegal practices in the stock market. In stock data analysis, important issues that are often considered as research topics include stock price prediction and detection of stock price manipulation.

For the problem of stock price prediction, there are various data analysis techniques, algorithms based on machine learning, or deep learning models that have been proposed to address this problem. In this paper, our aim is to analyze and forecast the VNIndex short-time data, which represents the entire Vietnamese stock market with high-frequency updated data, using LSTM models (Hochreiter, 1997). Long Short-Term Memory networks (LSTM) are an extension of the recurrent neural networks (RNN) that can be effectively used for prediction of time series.

An autoencoder-based LSTM (LSTM-AE) was used as the unsupervised learning part for anomaly detection in normal cases, synthesized manipulation from VNIndex. The idea can be extended to use the ensemble learning model with a combination of both supervised methods to better understand anomalies. An ensemble stacking classifier model combining two classification methods SVM and ANN based on labeled data and known anomaly patterns were also implemented for anomaly detection.

2. LITERATURE REVIEW

Prediction of VNIndex using LSTM algorithm

Stock market have attracted more attention in the modern economy of Vietnam. Stock price predictions have a significant impact on decision making in stock investment and become an interesting research topic for researchers, investors and government administrators. LSTM algorithms, which function on computational gates to handle data and keep more useful information, have outperformed the regular RNN networks in stock price forecasting. Trung-Tuan Nguyen (2022) has investigated the Stacked LSTM models to optimize the price forecasting and verified them with daily stock prices from three stock codes AAPL (in the US), AAA and VCB (in Vietnam). A more comprehensive study of 30 stock codes from Ho Chi Minh City stock exchange, Vietnam (Phuoc Tran, 2024) using LSTM algorithm and technical analysis indicators results in good forecasting level for most data of the stock codes studied.

Detection of stock anomaly/stock manipulation using machine learning

In the machine learning approach, the techniques can be divided into two types: Supervised learning and unsupervised learning. Supervised learning based on prior information can achieve high accuracy. However, some recent studies have focused on some well-known manipulation patterns such as pump-and-dump, because of labeling.

Cao (2015) implemented an adaptive hidden Markov model with anomaly states (AHMMAS) for stock price manipulation detection of stocks from NASDAQ and the London Stock Exchange. Wang et al. (2019) used RNN for stock manipulation detection which LSTM itself is also a RNN model for stock manipulation. Leangarun (2021) used the LSTM-GANs and LSTM-AE to detect manipulation without prior knowledge but with low accuracy. Chullamonthon (2023) utilized the idea of an ensemble of both supervised and unsupervised learning with deep neural network structures and have achieved higher accuracy for both synthesized manipulation cases and real manipulation cases.

Synthesized study of stock manipulation

The knowledge and proposed degrees of pump-and-dump manipulations with more details based on previous studies by Kamps (2018) and Chullamonthon (2023). This would be used for preparing synthesized manipulation data used in model training for stock price manipulation detection as performed previous research (Cao, 2012).

3. METHODOLOGY

LSTM is a deep learning neural network architecture that is commonly used in time series forecasting problems. With the ability to learn and remember long-term information, LSTM can model complex relationships in data sequences and predict the future data with high accuracy.

This enables LSTM to predict the next values in the data series accurately and reliably. However, solving the problem of anomaly forecasting with LSTM is not simple, because anomalies can appear at any position in the data series and can be random. To extract the features for LSTM method from a time-series data effectively, we utilized the Wavelet Scattering Transform (WST) (Soro, 2018). WST is a multi-layer transformation that applies wavelet filters to a signal to decompose it into its frequency components and then uses a series of nonlinear transformations to compute the signal's features. A recent work by Jagadesh (2024) has shown the impact of wavelet transform in improving data quality for financial data from stock market including data cleaning and noise reducing to provide high-quality input data for the forecasting and anomaly-detection models.
In addition, supervised machine learning algorithms such as ANN and SVM are also considered as one of the most popular data mining techniques used in anomaly detection. Although they have some limitations, such as dependence on labeled data and known anomaly patterns, as well as considering factors that depend on the available data, the idea can be extended to use ensemble learning models with a combination of both supervised and unsupervised methods to better understand anomalies.

4. PREDICTION OF VNINDEX USING UNIVARIATE LSTM

4.1. Data preprocessing

The Vietnam stock data set is provided by BK Fintech Database. This database contains historical stock data including all information from 08-03-2021 to 12-01-2021 of all stock codes in Vietnam updated every 5 seconds. Below is an illustration of the data set of FPT Joint Stock Company.

	SYMBOL	DATE	OPEN	HIGH	LOW	CLOSE	VOLUME	OPENINT
0	FPT	2023-08-03 09:15:05	82.8	None	None	None	11300.0	0.0
1	FPT	2023-08-03 09:15:08	82.7	None	None	None	100.0	0.0
2	FPT	2023-08-03 09:15:15	82.7	None	None	None	100.0	0.0
3	FPT	2023-08-03 09:15:59	82.7	None	None	None	300.0	0.0
4	FPT	2023-08-03 09:16:10	82.7	None	None	None	600.0	0.0
...	...	...	...	...	...	...	...	...
90331	FPT	2023-12-01 14:29:35	91.8	None	None	None	1000.0	0.0
90332	FPT	2023-12-01 14:29:40	91.9	None	None	None	3000.0	0.0
90333	FPT	2023-12-01 14:29:45	91.7	None	None	None	9500.0	0.0
90334	FPT	2023-12-01 14:29:50	91.9	None	None	None	200.0	0.0
90335	FPT	2023-12-01 14:45:02	92.4	None	None	None	158600.0	0.0

[90336 rows x 8 columns]

Fig. 1. Stock historical data set of FPT from 08-03-2021 to 12-01-2021

4.2. Prediction results

The univariate LSTM model (Moghar, 2020) was trained on 20000 samples of FPT stock price up to 11-12-2021 including 70% of training data and 30% of validation data. The model was then utilized to test the prediction for the whole dataset. Figure 2 illustrates the comparison between the actual and predicted stock price data using an LSTM model.

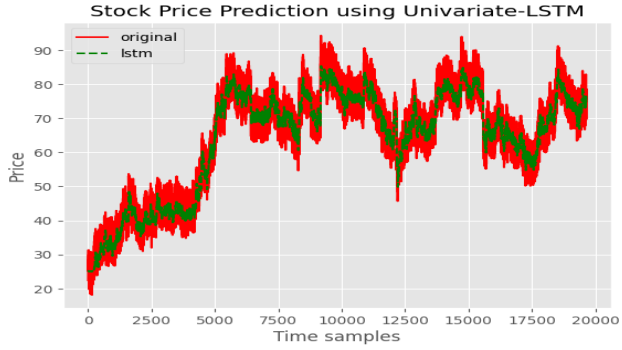


Fig. 2. Comparison between the actual and predicted stock price data using an LSTM model

The "original" line, depicted in red, represents the true number of infection cases at each time point in the data. The 'lstm' line, shown in green, represents the LSTM model's forecasts. We choose the number of time steps ahead to be predicted as 360 samples (30 minutes). The LSTM model can capture the underlying trends and dynamics of the evaluated data. We select the mean absolute percentage error (MAPE) and Coefficient of Determination (R-square) for the forecasting accuracy measures.

Time (samples)	MAPE	R-square
360	0.0451	0.9528

5. ANOMALY DETECTION USING LSTM-AUTOENCODER AND ENSEMBLE STACKING

5.1. LSTM - Auto encoder (LSTM-AE) for anomaly detection

Based on the LSTM method presented in the above section, we will construct the LSTM-AE (Faraz, 2020) with a combination of some basic layers such as LSTM layer, dropout layer, ReLU layer and the last layer will use regression layer. In addition, the model needs some settings such as optimization method (currently using Adam), maximum number of steps, minibatch size, etc. With three datasets (training, validation and testing) containing the features attracted using wavelet scattering transform (WST), we take them as input for the model to train and give the predicted values. The mean absolute error (MAE) then was used to represent the difference between the actual data and the predicted value.

The MAE value of the validation feature set is used to generate a threshold value, which is the value that determines the output for the data series. The upper quartile of the MAE value of the validation set plus 1.5 times of its interquartile range is used as the threshold in this work. This is a conservative approach and will minimize the risk of missing real anomalies.

(a) (b)

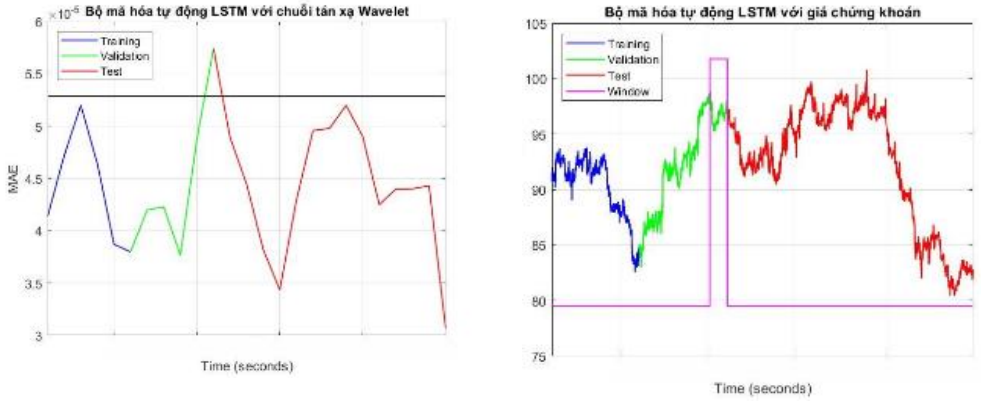


Fig. 3. LSTM-AE model trained with normal data (FPT stock price): (a)MAE error; (b) Time series data and the window outlining the anomaly

The LSTM-AE model was trained with normal data (FPT stock price), and it tried to reconstruct the normal input data to be the same as the original data. We need to measure how far the reconstruction input is from the original using the construction error (MAE). In price manipulation detection, if any time series data tested with this model has a small reconstruction error or short period when the MAE value is higher than the given threshold, that time series tends to be normal.

5.2. Synthesized manipulation data

The manipulation levels (low, median, high) will be used to synthesize the manipulation data (Chullamonthon, 2023). In general, we define T_p as the pump length and T_d as the dump length. Figure 4 shows the scheme for the increase in price during the pump phase $0 < t < T_p$ and the decrease in the dump phase $(T_p < t < T_p + T_d)$.

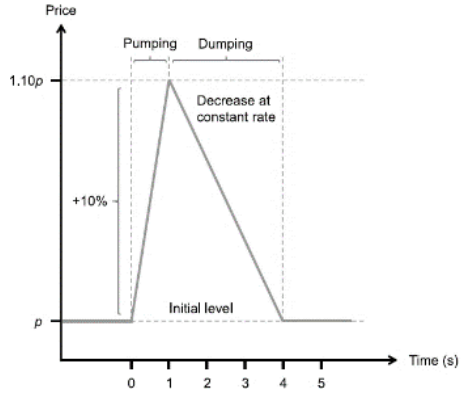


Fig. 4. Typical pump-and-dump scheme

The pumping and dumping injection in the price is shown in the equation (1).

$$P_{injected,t} = P_0 + \gamma_p \Delta P_t \quad (1)$$

$$\Delta P_t = \begin{cases} K_p t P_0, 0 < t < T_p \\ \left(1 - K_d(t - T_d)\right) P_0, T_p < t < T_p + T_d \end{cases}$$

where: P_0 is the stock price at the initial time, ΔPt is the price difference from the initial time to t, $P_{injected,t}$ is the stock price after manipulation injection at time t. γ_p is the price change ratio, which determines how much the price will change in each degree of manipulation. The value of γ_p can be 0.03 to 0.10 (3% to 10%).

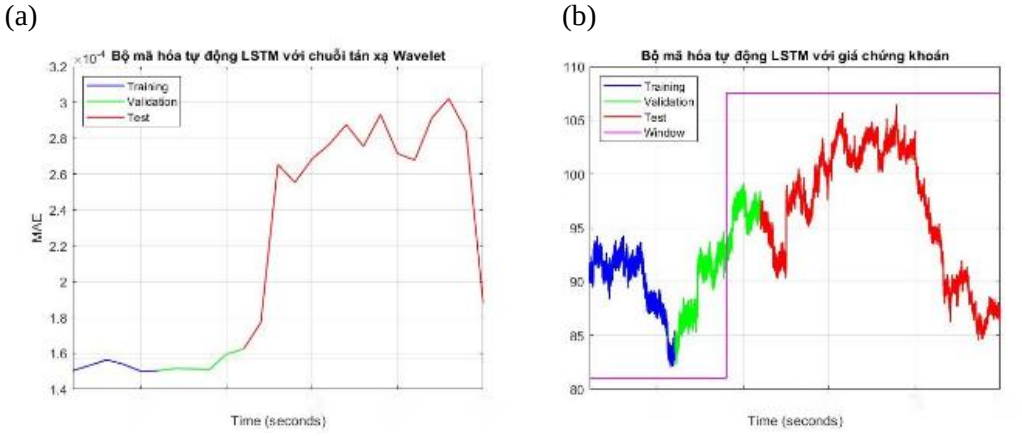


Fig. 5. LSTM-AE model trained with synthesized data (FPT stock price): (a)MAE error; (b) Time series data and the window outlining the anomaly

In case of synthesized manipulation, the model has a large reconstruction error, that time series tends to be a manipulated one. The unsupervised model used an autoencoder to learn and treat unrecognized behaviors as anomalies (manipulations).

5.3. Ensemble stacking model for anomaly detection

The ensemble stacking method combined supervised based models for classification using classified labels from unsupervised LSTM-AE method to gain the accuracy of anomaly classification (Chullamonthon, 2023). We include support vector machine (SVM), Logistic regression (LP) and Artificial neuron network (ANN-MLP) as based models because of its high robustness. We have tested on 8 stock codes in Vietnam, including HPG, DIG, NVL, VIC, HSG, VNM, BCG, FPT and the synthetic dataset based on these 8 stock codes with the length of each data series being 90,000 data points. These codes are labeled based on the results of the algorithm using LSTM-AE with 12 abnormal codes (synthesized) and 8 normal codes. The accuracy of the base model and metal model using k-fold cross validation. The experimental results show that the ensemble model outperforms the individual models.

Models	SVM	Logistic regression	MLP	Ensemble stacking
Accuracy score	0.9094	0.8182	0.9089	0.9241

Figure 6 is the confusion matrix representing the classification results on the test set with 6 stock codes (3 normal, 3 unnormal). Only one out of 6 codes was mistakenly classified.

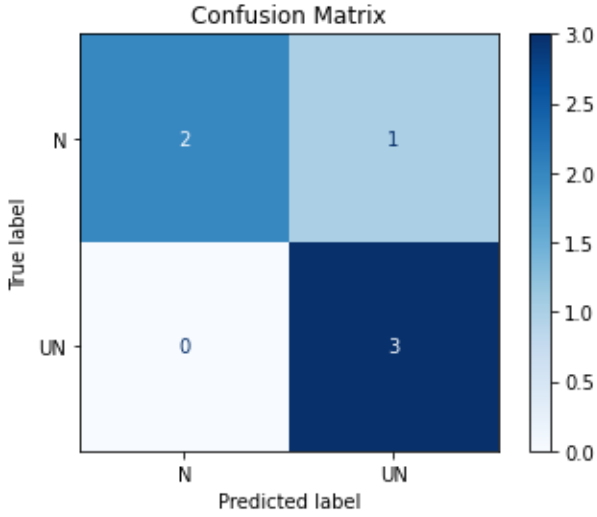


Fig. 6. Confusion matrix representing the classification results.

6. CONCLUSION AND DISCUSSION

In this paper, we proposed a framework for detecting stock price manipulation. We gathered short time trading data in Vietnam in 2021 and performed a stock price prediction using univariate LSTM. The main contribution was to use a normal dataset to train an unsupervised learning model LSTM-AE that could learn normal trading behaviors and identify outliers as anomalies. We also synthesized difference degrees of pump-and-dump manipulation and injected them into our normal dataset for training the unsupervised learning model LSTM-AE model and got the high accuracy of anomaly detection. The outcomes of the real data and synthesized data analysis using LSTM-AE models for 10 stock codes from VNINDEX then utilized for training of a supervised learning ensemble stacking model. The proposed method can accurately classify the abnormal ones from unlabeled data set. For future work, proposed method can be applied in real time detection of stock manipulation in Vietnam stock market. The wide range of factors such as investor behavior, stock company performance, geopolitical events and economic indicators that may impact the detection model should be considered for the future work.

ACKNOWLEDGMENT

The authors gratefully acknowledge Vietstock for providing online access to the VNIndex stock price and Dr. Tran Ngoc Thang for the suggestions to improve the manuscript. We also want to thank you Bkfintech research center and the School of Applied Mathematics and Informatics, Hanoi University of Science and Technology for their support to the work.

REFERENCES

Soro, Bedionita, and Chaewoo Lee (2018). Performance comparison of indoor fingerprinting techniques based on artificial neural network". *TENCON 2018-2018 IEEE Region 10 Conference*. IEEE.

Cao, Longbing, Yuming Ou, and S. Yu Philip (2011). Coupled behavior analysis with applications". *IEEE Transactions on Knowledge and Data Engineering* 24.8: 1378-1392.

Cao, Yi, et al. (2014). Adaptive hidden Markov model with anomaly states for price manipulation detection". *IEEE Transactions on Neural Networks and Learning Systems* 26.2: 318-330.

Chullamonthon, Phakhawat, & Poj Tangamchit (2023). "Ensemble of supervised and unsupervised deep neural networks for stock price manipulation detection". *Expert Systems with Applications* 220: 119698.

Faraz, Mehrnaz, Hamid Khaloozadeh, and Milad Abbasi (2020). "Stock market prediction-by-prediction based on autoencoder long short-term memory networks." *28th Iranian conference on electrical engineering (ICEE)*. IEEE, 2020.Hirsa, A. (2013). *Computational Methods in Finance.* Taylor & Francis Group, LLC: 351-359.

Hochreiter, Sepp, and Jürgen Schmidhuber (1997). Long short-term memory. Neural computation 9.8: 1735-1780.

Jagadesh, B. N., et al. (2024) "Enhanced stock market forecasting using dandelion optimization-driven 3D-CNN-GRU classification." *Scientific Reports* 14.1: 20908.

Kamps, Josh, and Bennett Kleinberg (2018). "To the moon: defining and detecting cryptocurrency pump-and-dumps". *Crime Science* 7.1: 1-18.

Leangarun, Teema, Poj Tangamchit, and Suttipong Thajchayapong. "Stock price manipulation detection using deep unsupervised learning: The case of Thailand." *IEEE Access* 9 (2021): 106824-106838.

Moghar, Adil, and Mhamed Hamiche (2020). "Stock market prediction using LSTM recurrent neural network". *Procedia Computer Science* 170: 1168-1173.

Phuoc, Tran, et al. (2024). "Applying machine learning algorithms to predict the stock price trend in the stock market–The case of Vietnam." *Humanities and Social Sciences Communications* 11.1: 1-18.

Tuan, Nguyen Trung, Thu Hang Nguyen, and Thanh Thi Hien Duong (2022). "Stock Price Prediction in Vietnam Using Stacked LSTM." *International Conference on Intelligence of Things*. Cham: Springer International Publishing.

Thu-Trang, Pham. (2022). "Stock Price Analysis for price prediction and anomaly detect" *Undergraduate student graduation thesis, HUST*.

“Vietstock.” Vietstock. [Online]. Available: https://vietstock.vn/

Wang, Qili, et al. (2019). "Enhancing intraday stock price manipulation detection by leveraging recurrent neural networks with ensemble learning". *Neurocomputing* 347: 46-58.