Current Journal of Applied Science and Technology

Volume 43, Issue 8, Page 1-13, 2024; Article no.CJAST.120766
ISSN: 2457-1024
(Past name: British Journal of Applied Science & Technology, Past ISSN: 2231-0843, NLM ID: 101664541)

A Novel Approach to the Moving Average Distance and Isolation Forest Method for Forecasting Financial Time Series Data

Agus Sihabuddin [a*], Nur Rokhman [a], Mohammad Edi Wibowo [a] and Abdul Karim [b]

[a] Department of Computer Science and Electronics, University of Gadjah Mada, Yogyakarta, Indonesia.
[b] College of Science and Technology, Korea University, Seoul, South Korea.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: https://doi.org/10.9734/cjast/2024/v43i84416

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/120766

Original Research Article

Received: 26/05/2024
Accepted: 28/07/2024
Published: 31/07/2024

ABSTRACT

Aims: The main aim of this research is to propose a new approach to financial time series data feature extraction using Moving Average Distance combined with Isolation Forest.
Study Design: Building a feature extraction using Moving Average Distance and Isolation Forest for time series data.
Methodology: We propose a Moving Average Distance to calculate the distance of the recent price to a moving average as input for the Isolation Forest algorithm. The distance measurement used in

**Corresponding author: E-mail: a_sihabudin@ugm.ac.id;*

Cite as: *Sihabuddin, Agus, Nur Rokhman, Mohammad Edi Wibowo, and Abdul Karim. 2024. "A Novel Approach to the Moving Average Distance and Isolation Forest Method for Forecasting Financial Time Series Data". Current Journal of Applied Science and Technology 43 (8):1-13. https://doi.org/10.9734/cjast/2024/v43i84416.*

this research is measured in daily periods of 2,3,4,5,10, and 20. This period variation is used to observe the effect of different short, medium, and long terms on the forecasting accuracy. This moving average distance and isolation forest combination enrich the feature used as input for LSTM as a forecasting algorithm.
Results: The daily S&P 500 and SSE index are used as datasets to test the proposed method. The research results showed that the proposed method outperformed the accuracy of previous research. The 3-daily period of moving average distance was the best parameter for the model and gave the best accuracy performance measure for both datasets; we found that the more extended period than that tends to reduce the accuracy.
Conclusion: Our experimental results showed that the proposed method improves the ability of LSTM as a time series forecasting algorithm and outperforms the previous research results.

Keywords: Moving average distance; isolation forest; abnormal detection; time series; forecasting; LSTM.

1. INTRODUCTION

Anomalies or outliers represent a few observations that show a significant deviation from the majority, often attributed to measurement inaccuracies or the absence of relevant covariates. In contrast to noise, classification, or attribute error, outliers include inconsistent data from natural variations or processes [1]. Outlier data can distort analysis results because it can affect centrality measures, increase variability, violate statistical assumptions, and affect the results of hypothesis tests and predictive models. Therefore, it is essential to identify and handle outlier data carefully. Two main strategies can be employed to tackle the outliers: deleting the outlier data or adjusting the methodology applied in the analysis [2].

Outlier detection has been used in various domains [3], including but not limited to fraud detection [4,5,6], weather forecasting [7,8,9], and financial time series [10-15].

Two prevalent approaches for identifying outliers are using descriptive (Z-Score [16], Range [17], Studentized Residuals [18], Minimum Covariance Determinant (MCD) [19], Cook's Distance [20], Mahalanobis Distance [21], and Local Outlier Factor (LOF) [19,22,23]) and machine learning (ML) clustering (k-Means [24], k-Means++ [25], Isolation Forest (IF) [26], and One-Class SVM [19,22]. One of the prominent techniques in the ML field for identifying and eliminating anomalies is known as the IF [1].

The IF can ascertain the anomaly scores by aggregating specific isolation trees [27]. The advantage of employing the IF is its ability to enhance computational effectiveness in high-dimensional datasets [28].

Nonetheless, it is crucial to consider the numerous constraints of the IF algorithm. According to [29], the IF algorithm has several shortcomings, including the fact that IF needs to be specifically designed to handle temporal correlation, which often exists in time series data. Furthermore, these algorithms necessitate an inherent mechanism to handle non-stationary data, a requirement that, if not met, could result in erroneous outlier identification when the data displays noteworthy trends or seasonal variations.

The Moving Average Distance (MAD) algorithm is a method to calculate the percentage distance of a price from a moving average. This kind of technical indicator is commonly used in financial research, such as prediction [30,31], market timing [32], and instrument returns [33]. We proposed MAD as a solution to overcome some of the shortcomings of IF in the context of financial time series data analysis. MAD is inherently designed to handle the temporal correlations often appearing in the time series data that IF suffers from. By using moving averages as a reference, MAD can effectively capture trends and seasonal variations in data, thereby reducing the risk of incorrectly identifying outliers in non-stationary data. A simple MAD makes it easier to detect more contextual anomalies and is sensitive to minor changes in the relationship between price and its moving average. This makes MAD more suitable for financial market technical analysis than the common IF. In addition, MAD does not require a training process and can be directly applied to data, making it more computationally efficient

and more accessible to implement for real-time analysis on financial time series datasets.

This study proposes a new approach to MAD as a distance feature enrichment for IF for the LSTM forecasting algorithm. The S&P500 index in the United States and the SSE index in China daily data are used to test the proposed method with evaluation metrics are R-squared (R^2), Root Mean Square Error (RMSE), Mean absolute percentage error (MAPE), and Mean Absolute Error (MAE).

2. METHODOLOGY

2.1 Moving Average Distance

Moving average distance is a price distance from a moving average used to determine whether the price is above or below its moving average. The commonly used moving average calculation is a simple moving average, with the other standard methods being weighted moving average, exponential moving average, and smoothed moving average. The distance is commonly expressed as a percentage of the moving average value, as in Eq. 1 and Eq. 2 [30].

$$MAD = \frac{(close - MA)}{MA} \quad (1)$$

$$\text{with } MA = \frac{\sum_1^n p_i}{n} \quad (2)$$

And p_i is the closing price, and n is the number of periods. This indicator is often combined with other MADs, and different periods and moving average methods are used.

2.2 Isolation Forest

The Isolation Forest (IF) is a collection of binary trees, called Isolation Trees, designed to isolate data points [28]. The algorithm generates individual isolation trees that merge into an ensemble method, the IF. The tree creation depends on the decisions determined by the data set format [28]. The IF functions optimally with massive datasets as it has a linear time complexity and low memory overhead [29]. So, the IF technique is an unsupervised approach to outlier detection from a collective-based method, where an isolation score is calculated for every data point [30]. Briefly, the distribution is split multiple times through IFs at random domain values, and then the number of splits required to isolate each point is counted. Points that require less splitting are more likely to be outliers.

The outlier score is determined by the number of necessary splits or output functions from numerous repetitions of this process [31].
To determine how an instance is unique compared to other cases based on their respective path lengths, it is essential to calculate the outlier score mathematically represented by Eq. (3), and Eq. (4) is used to evaluate the isolated trees' average path length [28,32,33].

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \quad (3)$$

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \quad (4)$$

with $s(x, n)$is an anomaly score; n is the size of the dataset; $E(h(x))$ is the average path length of the instance x over a tree collection and $c(n)$ is an average path length of an unsuccessful search in a binary search tree given n samples. $H(n-1)$ is a harmonic number and can be approximated by $ln(n-1) + 0.5772156649$ [34].

2.3 Long Short-Term Memory

LSTM networks are a specialized variant of recurrent neural networks (RNN) meticulously engineered to address sequence prediction challenges [35]. Their distinctive architecture, which facilitates the retention of patterns over extended durations, renders LSTMs exceptionally proficient for time series modeling.

LSTM networks consist of memory cells that are regulated by three gates: forget gate (ft) input gate (it), and output gates (ot). These gates determine how information flows through the memory cells as in Eq. (5),(6),(7),(8) and (9).

$$\mathrm{ft} = \sigma(\mathrm{Wf} \cdot [\mathrm{ht}-1, \mathrm{xt}] + \mathrm{bf}) \quad (5)$$

$$\mathrm{it} = \sigma(\mathrm{Wi} \cdot [\mathrm{ht}-1, \mathrm{xt}] + \mathrm{bi}) \quad (6)$$

$$\tilde{\mathrm{C}}\mathrm{t} = \tanh(\mathrm{WC} \cdot [\mathrm{ht}-1, \mathrm{xt}] + \mathrm{bC}) \quad (7)$$

$$\mathrm{Ct} = \mathrm{ft} \times \mathrm{Ct}-1 + \mathrm{it} \times \tilde{\mathrm{C}}\mathrm{t} \quad (8)$$

$$\mathrm{ot} = \sigma(\mathrm{Wo}[\mathrm{ht}-1, \mathrm{xt}] + \mathrm{bo}) \quad (9)$$

$$\mathrm{ht} = \mathrm{ot} \times \tanh(\mathrm{Ct}) \quad (10)$$

with σ is the sigmoid function, W, and b are the weight matrices and biases for each gate, respectively, xt is the input at time t, and ht is the output.

2.4 Proposed Method

In this study, we propose using MAD as a feature of IF in a forecasting model with LSTM Algorithm. There are two main contributions to this research:

1. The MAD as distance feature for IF in anomaly detection for forecasting LSTM algorithm input
2. The effects of MAD periods, 2,3,4,5,10, and 20 for the forecasting accuracy results.

The algorithm started by inputting the time series dataset, namely the S&P500 and SSE daily index. The MAD was calculated with 2,3,4,5,10, and 20 periods for both datasets. The price and MAD value are processed with the IF anomaly detection algorithm to calculate the IF anomaly value index. The moving average calculation method used in this research is a simple moving average.

Data is split into training and testing data. An LSTM forecasting model is generated and trained using the training data to get the best model for every period of MAD. The best LSTM-generated model is used to test data to evaluate the model's accuracy using different forecasting key metrics such as R^2, MAPE, RMSE, and MAE. The forecasting period is one step ahead or a one-time horizon. The whole process is presented in Algorithm 1.

2.5 Evaluation Metrics

The MAE is a metric used to evaluate the regression method. It calculates the mean of the predicted errors over all instances to give the final score. It assesses the variation between the expected value of an instance and its actual value [36,37]. This is simple to measure and less sensitive to outlying values [38]. Equation (11) is a description of the metrics used in this research work [36,39].

$$MAE = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda\left(x_i\right)\right)}{n} \quad (11)$$

with y_i is the data ground truth value for x_i, $\lambda\ (x_i)$ is the predicted value for a data x_i , x_i is the number of data.

The R^2 or coefficient of determination is a statistical measure quantifying uncertainty from 0 to 1. A value of 1 indicates a strong correlation between estimated and measured values [40]. R^2 is given by Equation (12) [41].

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_i - y_m\right)\left(\lambda\left(x_i\right) - \lambda\left(x_m\right)\right)}{\sqrt{\left(\sum_{i=1}^{n}\left(y_i - y_m\right)^2\right)\left(\sum_{i=1}^{n}\left(\lambda\left(x_i\right) - \lambda\left(x_m\right)\right)^2\right)}} \quad (12)$$

where y_m and $\lambda\ (x_m)$ are the mean of the actual and predicted values [42-44].

RMSE is the root mean square error of the prediction versus the observation. RMSE is shown by Equation (13) [41].

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \lambda(x_i))^2}{n}} \quad (13)$$

Algorithm 1. Hybrid moving average distance and IF on forecasting method

Input: Data Set
Output: LSTM Model and the accuracy Process: 1. Input data set time series. 2. Calculate the MAD 3,5,10 and 20 period 3. Anomaly detection from the dataset using IF with MAD feature. 4. Split data into train and test sets. 5. Generate an LSTM forecasting method. 6. Train the LSTM forecasting method. 7. Evaluate the method accuracy using R^2, MAPE, RMSE, and MAE.

3. RESULTS AND DISCUSSION

Two datasets were used in this research: daily stock prices of the S&P 500 and SSE Index. The S&P 500 and SSE Index spanning December 31, 2012 to December 31, 2022 [45-47]. These data were chosen to demonstrate the financial dataset as in [48]. The data contains open, high, low, close, and volume. The data used for the algorithm is close. The data was downloaded from Yahoo Finance API, reliable data commonly used for financial research. The preprocessing used in this dataset is only null data cleaning, in which the amount of null data is very small. The S&P500 and SSE index data count are 2519 and 2428, respectively. The data split for both is 80% data for training and 20% for testing, as in Table 1.

The MAD period of 2,3,4,5,10,20 is used to see the different MAD periods to the behavior of the accuracy of the LSTM forecasting algorithm. These periods of 2, 3, and 4 are used to cope with short-term daily data trading of less than a week. The 5,10,20 periods are used to deal with weekly, bi-weekly, and monthly patterns, respectively.

Fig. 1 illustrates the segmentation of the S&P 500 index dataset into two clearly defined sections: the training dataset (Train) and the testing dataset (Test). The temporal extent along the X-axis of the chronology ranges from 2013 to 2022, whereas the Y-axis illustrates the S&P 500 stock prices denoted in US dollars.

The training dataset, delineated by the blue line, encompasses the timeframe spanning from 2012 to the beginning of 2021 and was utilized to train the machine learning algorithm. On the contrary, the test dataset, indicated by the orange line, extends from the beginning of 2021 to 2022 and was utilized for assessing the effectiveness of the trained model.

This partitioning strategy is implemented to mitigate overfitting and ensure the robust performance of the model on unseen data.

Table 1. Dataset descriptive

Dataset	Time Range	Time Frame	Count	Data Split
S&P 500	Dec. 31, 2012 to Dec. 31, 2022	Daily	2519	80:20
SSE Index	Dec. 31, 2012 to Dec. 31, 2022	Daily	2428	80:20

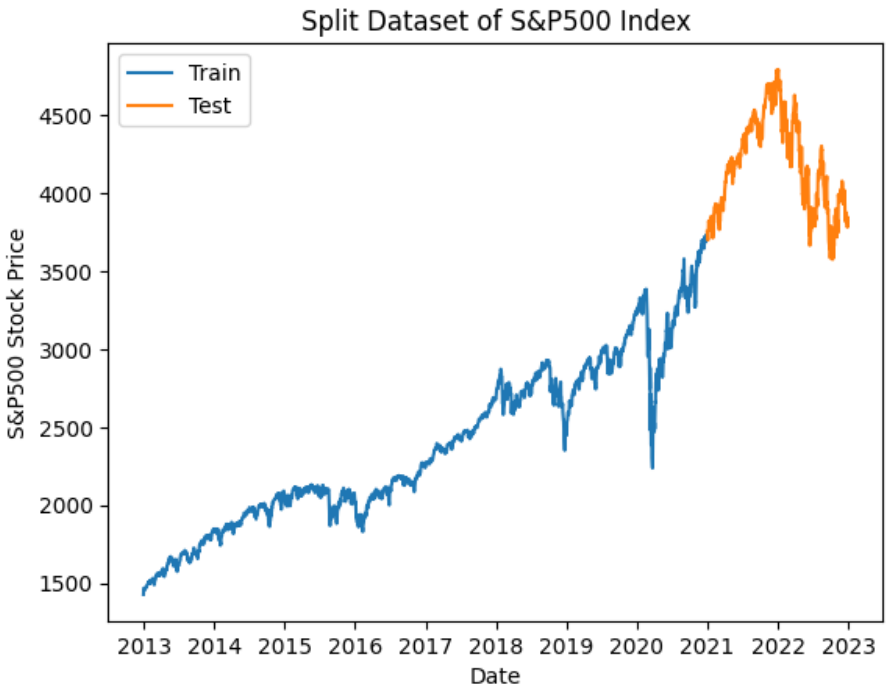

Fig. 1. The S&P index price from 2012 until 2022

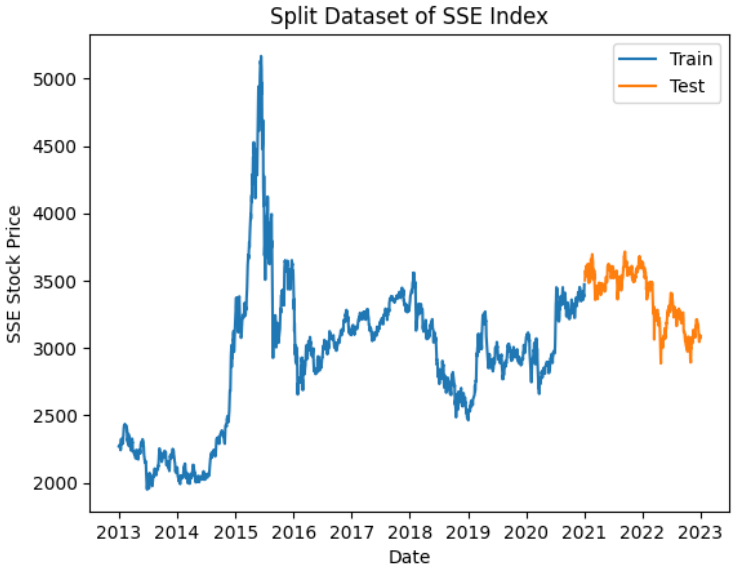

Fig. 2. The SSE index price from 2012 until 2022

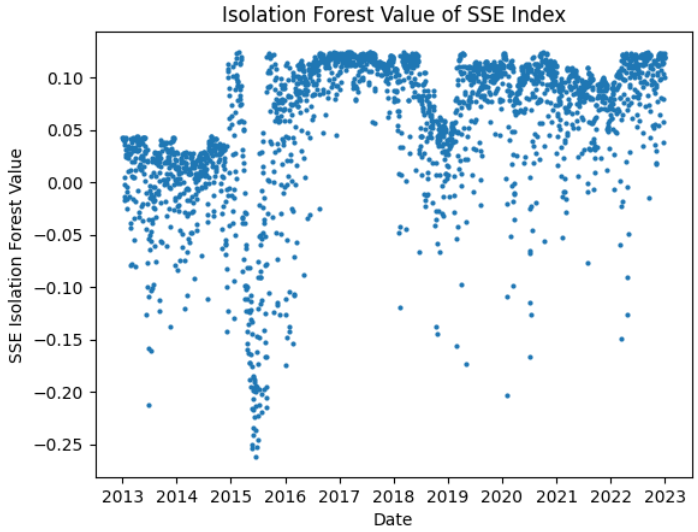

Fig. 3. The isolation forest graph of SSE IndexSSE index

Fig. 2 shows a graph dividing the SSE Index dataset into training data (Train) and testing data (Test). The horizontal axis represents the timeframe spanning from 2012 to 2022, whereas the vertical axis illustrates the value of SSE shares. The blue line illustrates the training dataset spanning from 2013 until the conclusion of 2020, while the orange line denotes the testing dataset covering the timeframe from the onset of 2021 to 2022.

The MAD calculation is applied for the training and testing data using the predetermined daily period. After this, the IF takes it as input and results in the anomaly score. Fig. 3. shows the results of applying the Isolation Forest method to the MAD of a 3-daily period of the SSE index data from 2012 to 2022.

The Lower (more negative) values indicate a greater likelihood that the data point is an anomaly. Points below the Y-axis indicate significant anomalies, as seen in 2015 and 2016. Most points are 0 to 0.1, indicating that the algorithm considers most data normal. This figure helps identify periods with significant anomalies in the SSE index data.

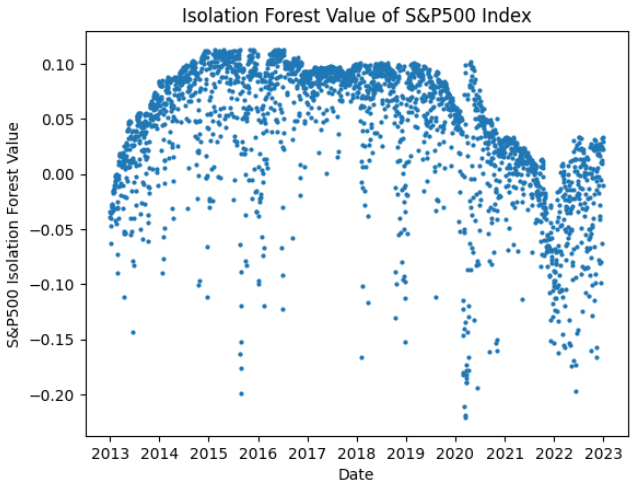

Fig. 4. The Isolation Forest graph of the S&P 500 index

Fig. 5. Model loss convergency for the SSE index

Fig. 4. depicts outcomes from applying the IF algorithm on the 3-period of MAD of the S&P500 index data from 2012 to 2022. The temporal dimension is represented on the horizontal axis (X), whereas the anomaly score generated algorithmically is displayed on the vertical axis (Y). High positive values indicate average data, while low negative values indicate anomalies. The 2012-2017 period tends to be expected, 2018-2020 shows an increase in anomalies, and 2020-2023 has large fluctuations with many anomalies, possibly related to significant COVID-19 pandemic events.

The time series forecasting algorithm, LSTM, is used to process the result of MAD+IF and the close price of data with the MSE loss function, ADAM optimizer, and 20 epoch.
The segregation proposed aims to optimize the training procedure of the predictive model by utilizing a distinct training dataset and evaluating the model's accuracy using a separate testing dataset. This approach guarantees the model's capacity to generalize effectively when encountering novel, unobserved data.

The Model Loss SSE Index graph in Fig. 5 compares Training Loss and Validation Loss during the ML model training process. Training Loss starts from a high value of around 0.005 and drops drastically in the first few epochs, while Validation Loss starts from a lower value and is relatively stable. Both lines converge after the 2.5th epoch, indicating that the model learns well without significant overfitting. The consistent decrease in loss and the final value close to zero indicates that the model accurately predicts the SSE Index. Overall, this graph depicts an effective learning process, with the model successfully reducing its prediction error

consistently on both data sets, achieving stable, good performance and not overfitting.

Fig. 6 shows changes in loss values during model training on S&P500 Index data, with two curves: Training Loss (blue line) and Validation Loss (orange line). The X-axis illustrates the quantity of epochs, whereas the Y-axis represents the loss value computed through the Sum of Squared Errors (SSE). The loss value is initially considerably elevated during training yet undergoes a sharp decline within the initial epochs, ultimately reaching a modest level. This indicates proficient learning from the training dataset by the model, enabling effective generalization on novel data while avoiding overfitting.

Table 2 presents the evaluation results for forecasting the S&P 500 Index compared to previous research data with the same data split and the same data set [48]. Model numbers one until five are from previous research, and the numbers six until eleven are the proposed method with different MAD periods: 2,3,4,5,10, and 20. The first proposed model (model number 6) with the MAD 2-daily period with IF processed with the LSTM algorithm gives better accuracy parameters for MAPE, RMSE, and MAE with results of 0.05%, 32.81, and 24.98, respectively, compared to previous research. It gives a bit lower accuracy for R^2 compared to ARIMA, Holt's LES, and SVR but slightly higher than LSTM and GRU.

The S&P 500 best result is achieved by model number 6 by MAD 3-daly period with IF and LSTM with R^2 of 99.99%, MAPE of 0.05%, RMSE of 1.88, and MAE of 1.64, with all the accuracy metrics outperforming all other models (previous research and other proposed methods).

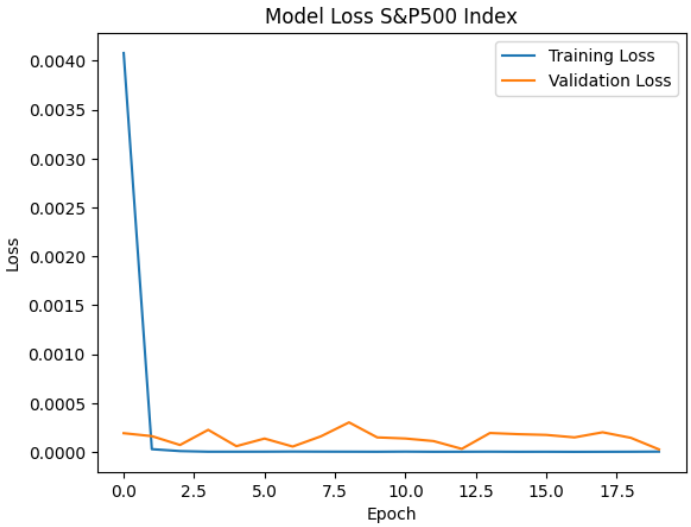

Fig. 6. Model loss convergency for the S&P 500 index

Table 2. The evaluation result for S&P 500 index

No	Models	R^2(%)	MAPE	RMSE	MAE
1	ARIMA [48]	99.04	1.06%	53.93	38.78
2	Holt's LES [48]	99.02	1.06%	54.48	38.62
3	SVR [48]	98.95	1.18%	56.31	43.16
4	LSTM [48]	94.70	2.71%	126.80	108.05
5	GRU [48]	98.16	1.60%	74.72	61.29
6	MAD (2)+IF+LSTM	98.82	0.57%	32.81	24.98
7	MAD (3)+IF+LSTM	99.99,	0.05%	1.88	1.64
8	MAD (4)+IF+LSTM	96.89	1.06%	53.30	46.18
9	MAD (5)+IF+LSTM	95.22	1.25%	66.08	54.58
10	MAD (10)+IF+LSTM	89.10	2.00%	99.79	86.24
11	MAD (20)+IF+LSTM	58.28	4.18%	195.26	179.17

Table 3. The evaluation results of the SSE index

No	Models	R^2(%)	MAPE	RMSE	MAE
1	ARIMA [48]	97.78	0.81%	36.25	26.34
2	Holt's LES [48]	97.72	0.81%	36.31	26.25
3	SVR [48]	97.81	0.80%	35.97	25.98
4	LSTM [48]	97.49	0.85%	38.54	27.72
5	GRU [48]	97.65	0.83%	37.27	27.19
6	MAD (2)+IF+LSTM	99.94	0.14%	4.81	4.70
7	MAD (3)+IF+LSTM	99.96	0.10%	4.04	3.28
8	MAD (4)+IF+LSTM	99.95	0.11%	4.44	3.78
9	MAD (5)+IF+LSTM	99.72	0.26%	10.66	8.78
10	MAD (10)+IF+LSTM	92.99	1.41%	53.01	47.93
11	MAD (20)+IF+LSTM	94.60	1.14%	46.55	37.99

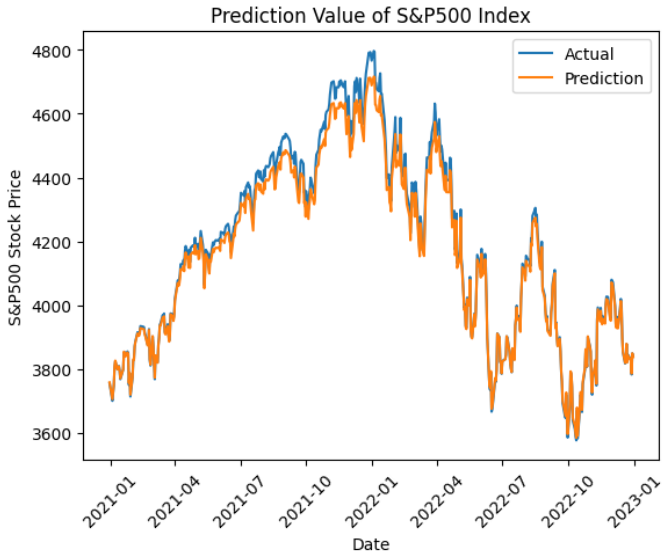

Fig. 7. The S&P accuracy performance graph

It can be seen that the proposed method model numbers 6,7,8,9,10, and 11 have a trend that the accuracy increases from model number 6 with MAD (2) to 7 with MAD (3) and then decreases the accuracy as the period of MAD period increases from 4,5,10 and 20. The decreasing rate increases as the period of daily MAD increases. The worst result is seen from the last model with 20-daily MAD continued with IF and LSTM with R^2 of 58.28%, MAPE of 4.18%%, RMSE of 195.26, and MAE of 179.17. It is the worst model, too, compared to other models presented in Table 2.

Table 3 has the same pattern as Table 2: the proposed method applied to the SSE index outperforms the previous research. Model numbers 7,8, and 9 outperform the R^2 from previous research, and models 6,7,8, and 9 outperform the MAPE, RMSE, and MAE values compared to previous research. It can be concluded that the proposed model outperforms the previous research. The best accuracy parameter is given by model number 7 with a MAD period of 3 with the value of R^2 of 99.96%, MAPE of 0.10%, RMSE of 4.04, and MAE of 3.28.

The accuracy pattern for a different MAD period is similar to that of the S&P 500 dataset in that the accuracy tends to increase for a MAD period of 2 to 3 and then decrease from a MAD of 3 to 4,5,10, and 20. The worst accuracy result is the last model with a 20-daily MAD period and IF plus LSTM, but the accuracy is close to other models in Table 2.

Both outperforming results from the S&P SSE Index can be obtained by the distance feature introduced by MAD for the IF algorithm. This

approach suggests that the MAD continued by the IF algorithm enriches the feature for the LSTM forecasting algorithm, and this approach solves the IF, which calculates the algorithm not by distance.

The envisaged chart of the S&P 500 Index (Fig. 7) demonstrates that the forecasting model (depicted by the orange line) exhibits a relatively close correspondence with the authentic data (represented by the blue line), albeit minor disparities are evident. A notable pattern observed entails a substantial surge in prices around mid-2021, succeeded by a sharp downturn in 2022, along with subsequent fluctuations until 2022. This visual representation is a valuable tool for analysts and investors in assessing the precision of predictive models and aiding in formulating investment decisions predicated on future price projections of the S&P 500 Index.

In summary, the proposed model of MAD for IF enrichment of distance feature for forecasting model has a promising result and outperforms the previous research. The shorter the MAD period, the better the forecast accuracy is than the more extended MAD period.

4. CONCLUSION

In this paper, we propose a MAD as a feature for the IF algorithm for anomaly detection in a forecasting model. The MAD enriches the distance feature for DTR based on IF. Our experimental results showed that the method improves the ability of LSTM as a forecasting algorithm and outperforms the previous research results. The smaller MAD period tends to have higher accuracy than the more extended periods of MAF.

Other moving average calculation methods, such as weighted, exponential, and smoothed moving averages, could be considered. A more extended daily period could be observed for long-term forecasts, which might give a better result. The more advanced IF algorithms, such as Extended IF, Generalied IF, or Deep IF, can be used to improve the basic IF algorithm's performance.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of manuscripts.

ACKNOWLEDGEMENTS

This work was partially supported by the Department of Computer Science and Electronics, Universitas Gadjah Mada, under the Publication Funding Year 2024.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

REFERENCES

1 Sihabuddin A, Rokhman N, Wahyudi EE. A machine learning approach on outlier removal for decision tree regression method. Ingénierie des Systèmes d'Information, vol. in press; 2024.

2 Basalamah S, Sihabuddin A. A Huber m-estimator algorithm and decision tree regression approach to improve the prediction performance of datasets with outlier. International Journal of Intelligent Engineering and Systems. 2024;17(1):1–9. DOI: 10.22266/ijies2024.0229.01.

3 Van NH, Van Thanh P, Tran DN, Tran DT. A new model of air quality prediction using lightweight machine learning. International Journal of Environmental Science and Technology. 2023, Mar;20(3):2983–2994. DOI: 10.1007/s13762-022-04185-w.

4 Huang Y, Liu W, Li S, Guo Y, Chen W. A novel unsupervised outlier detection algorithm based on mutual information and reduced spectral clustering. Electronics (Basel). 2023, Dec;12(23):4864. DOI: 10.3390/electronics12234864.

5 Mazarei A, Sousa R, Mendes-Moreira J, Molchanov S, Ferreira HM. Online boxplot derived outlier detection. Int J Data Sci Anal. 2024, May. DOI: 10.1007/s41060-024-00559-0.

6 Iqbal A, Amin R, Alsubaei FS, Alzahrani A. Anomaly detection in multivariate time series data using deep ensemble models. Plos One. 2024, Jun;19(6):e0303890. DOI: 10.1371/journal.pone.0303890.

7 Chen D, Lu CT, Kou Y, Chen F. On detecting spatial outliers. Geoinformatica. 2008, Dec;12(4):455–475. DOI: 10.1007/s10707-007-0038-8.

8 Amidi A, Hamm NAS, Meratnia N. Wireless sensor networks and fusion of contextual information for weather outlier detection. in

International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives. 2013;37–41.

9 Luo X, Heck B, Awange JL. Improving the estimation of zenith dry tropospheric delays using regional surface meteorological data. Advances in Space Research. 2013, Dec;52(12):2204–2214.
DOI: 10.1016/j.asr.2013.09.005.

10 Naidoo V, Du S. A deep learning method for the detection and compensation of outlier events in stock data. Electronics (Switzerland). 2002;11(21).
DOI: 10.3390/electronics11213465.

11 Shein WH, Ing NL, Fitrianto A. Stock market anomaly detection: Case study of China's securities market insider trading. In AIP Conference Proceedings; 2022.
DOI: 10.1063/5.0109178.

12 Akbar S, Saba T, Bahaj SA, Inshal M, Khan AR. Forecasting volatility in generalized autoregressive conditional heteroscedastic (GARCH) model with outliers. Journal of Advances in Information Technology. 2023;14(2):311–318.
DOI: 10.12720/jait.14.2.311-318.

13 Yaqoob T, Maqsood A. The potency of time series outliers in volatile models: An empirical analysis of fintech, and mineral resources. Resources Policy. 2024;89.
DOI: 10.1016/j.resourpol.2024.104666.

14 Mbiva SM, Correa FM. Machine learning to enhance the detection of terrorist financing and suspicious transactions in migrant remittances. Journal of Risk and Financial Management. 2024;17(5).
DOI: 10.3390/jrfm17050181.

15 Darné O, Levy-Rueff G, Pop A. The calibration of initial shocks in bank stress test scenarios: An outlier detection based approach. Econ Mode. 2024;136.
DOI: 10.1016/j.econmod.2024.106744.

16 Li Z, Xu R, Luo X, Cao X, Sun H. Short-term wind power prediction based on modal reconstruction and CNN-BiLSTM. Energy Reports. 2023, Dec;9:6449–6460.
DOI: 10.1016/j.egyr.2023.06.005.

17 Dobos D et al. A comparative study of anomaly detection methods for gross error detection problems. Comput Chem Eng. 2023, Jul;175:108263.
DOI: 10.1016/j.compchemeng.2023.108263.

18 Lim FP, Wong LL, Yap HK, Yow KS. Identifying outlier subjects in bioavailability trials using generalized studentized residuals. Sains Malays. 2023, May;52(5):1581–1593.
DOI: 10.17576/jsm-2023-5205-19.

19 Karasmanoglou A. ECG-based semi-supervised anomaly detection for early detection and monitoring of epileptic seizures. Int J Environ Res Public Health. 2023, Mar;20(6):5000.
DOI: 10.3390/ijerph20065000.

20 Dekker V, Schweikert K. A comparison of different data-driven procedures to determine the bunching window. Public Finance Review. 2021, Mar;49(2):262–293.
DOI: 10.1177/1091142121993055.

21 Sardashti A, Nazari J. A learning-based approach to fault detection and fault-tolerant control of permanent magnet DC motors. Journal of Engineering and Applied Science. 2023, Dec;70(1):109.
DOI: 10.1186/s44147-023-00279-5.

22 Kumar M, Saifi Z, Krishnananda SD. Decoding the physiological response of plants to stress using deep learning for forecasting crop loss due to abiotic, biotic, and climatic variables. Sci Rep. 2023;13(1):8598.
DOI: 10.1038/s41598-023-35285-3.

23 Huertas Celdrán A et al. Behavioral fingerprinting to detect ransomware in resource-constrained devices. Comput Secur. 2023, Dec;135:103510.
DOI: 10.1016/j.cose.2023.103510.

24 Ma J, Zhang H, Yang S, Jiang J, Li G. An improved robust sparse convex clustering. Tsinghua Sci Technol. 2023, Dec;28(6): 989–998.
DOI: 10.26599/TST.2022.9010046.

25 Liu X, Gong W, Shang L, Li X, Gong Z. Remote sensing image target detection and recognition based on YOLOv5. Remote Sens (Basel). 2023, Sep; 15(18):4459.
DOI: 10.3390/rs15184459.

26 Buck L et al. Anomaly detection in mixed high-dimensional molecular data. Bioinformatics. 2023, Aug;39(8).
DOI: 10.1093/bioinformatics/btad501.

27 Wang X, Ping W, Al-Shati AS. Numerical simulation of ozonation in hollow-fiber membranes for wastewater treatment. Eng Appl Artif Intell. 2023, Aug;123:106380.
DOI: 10.1016/j.engappai.2023.106380.

28 Heigl M, Anand KA, Urmann A, Fiala D, Schramm M, Hable R. On the improvement of the isolation forest

algorithm for outlier detection with streaming data. Electronics (Basel). 2021, Jun;10(13)1534.
DOI: 10.3390/electronics10131534.

29 Yang H, Li S, Tu L, Ma R, Chen Y. Unsupervised outlier detection mechanism for tea traceability data. IEEE Access. 2022;10:94818–94831.
DOI: 10.1109/ACCESS.2022.3204760.

30 Ejder U, Özel SA. A novel distance-based moving average model for improvement in the predictive accuracy of financial time series. Borsa Istanbul Review. 2024, Mar;24(2)376–397.
DOI: 10.1016/j.bir.2024.01.011.

31 Avramov D et al. Moving average distance as a predictor of equity returns we are grateful to moving average distance as a predictor of equity returns.
Available:https://ssrn.com/abstract=3111334

32 (Meni) Abudy M, Kaplanski G, Mugerman Y. Market timing with moving average distance: International evidence; 2023.

33 Alajbeg D, Bubas Z, Vasic D. Price distance to moving averages and subsequent returns; 2017. Available: http://ijecm.co.uk/

34 Barbariol T, Susto GA. TiWS-iForest: Isolation forest in weakly supervised and tiny ML scenarios. Inf Sci (N Y). 2022;610:126–143.
DOI: 10.1016/j.ins.2022.07.129.

35 Hochreiter Sepp, Schimidhuber Jurgen. Long short-term memory. Neural Comput. 1997;9(8):1735–1780.

36 Bao J, Adcock J, Li S, Jiang Y. Enhancing quality control of chip seal construction through machine learning-based analysis of surface macrotexture metrics. Lubricants. 2023;11(9).
DOI: 10.3390/lubricants11090409.

37 Mohy-Eddine M, Guezzaz A, Benkirane S, Azrour M, Farhaoui Y. An ensemble learning based intrusion detection model for industrial IoT security. Big Data Mining and Analytics. 2023;6(3).
DOI: 10.26599/BDMA.2022.9020032.

38 Liu Z, Zhao X, Tian Y, Tan J. Development of compositional-based models for prediction of heavy crude oil viscosity: Application in reservoir simulations. J Mol Liq. 2023;389.
DOI: 10.1016/j.molliq.2023.122918.

39 Gregg JT, Moore JH. STAR_outliers: A Python package that separates univariate outliers from non-normal distributions. BioData Min. 2023;16(1).
DOI: 10.1186/s13040-023-00342-0.

40 Călin AD, Coroiu AM, Mureşan HB. Analysis of preprocessing techniques for missing data in the prediction of sunflower yield in response to the effects of climate change. Applied Sciences (Switzerland). 2023;13(13).
DOI: 10.3390/app13137415.

41 Zheng H, Hu Q, Yang C, Mei Q, Wang P, Li K. Identification of spoofing ships from automatic identification system data via trajectory segmentation and isolation Forest. J Mar Sci Eng. 2023;11(8).
DOI: 10.3390/jmse11081516.

42 Oloyede A, Ozuomba S, Asuquo P, Olatomiwa L, Longe OM. Data-driven techniques for temperature data prediction: Big data analytics approach. Environ Monit Assess. 2023;195(2).
DOI: 10.1007/s10661-023-10961-z.

43 Tipu RK, Suman, Batra V. Enhancing prediction accuracy of workability and compressive strength of high-performance concrete through extended dataset and improved machine learning models. Asian Journal of Civil Engineering; 2023.
DOI: 10.1007/s42107-023-00768-1.

44 Shakeera S, Bala Naga Jyothi V, Venkataraman H. ML-based techniques for prediction of Ocean currents for underwater vehicles. 2023 11th International Symposium on Electronic Systems Devices and Computing, ESDC; 2023.
DOI: 10.1109/ESDC56251.2023.10149859.

45 Wudil YS, Imam A, Gondal MA, Ahmad UF, Al-Osta MA. Application of machine learning regressors in estimating the thermoelectric performance of Bi2Te3-based materials. Sens Actuators A Phys. 2023;351.
DOI: 10.1016/j.sna.2023.114193.

46 Rahman MM et al. Prospective methodologies in hybrid renewable energy systems for energy prediction using artificial neural networks. Sustainability. 2021, Feb;13(4):2393.
doi: 10.3390/su13042393.

47 Kassim NM et al. An Adaptive decision tree regression modeling for the output power of large-scale solar (LSS) farm forecasting. Sustainability. 2023;15(18):13521.
DOI: 10.3390/su151813521.

48 Jin S. A comparative analysis of traditional and machine learning methods in forecasting the stock markets of China and the US. IJACSA) International Journal of Advanced Computer Science and Applications. 2024; 15(4).
Available: www.ijacsa.thesai.org

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/120766