Applied Research and Smart Technology (ARSTech) Vol. 4 No. 2 (2023) 53–63

ARS Tech

Applied Research and Smart Technology (ARSTech)

The author(s) retain copyright

Journal homepage: https://journals2.ums.ac.id/index.php/arstech

ISSN 2722-9645

Research Article

Identification of stock market manipulation using a hybrid ensemble approach

Pearse Quinn[1*], Marinus Toman[1], Kevin Curran[1]

[1]*School of Computing, Engineering and Intelligent Systems, Ulster University, United Kingdom.*

[*]Corresponding author: Quinn-p43@ulster.ac.uk

Permalink (DOI): https://doi.org/10.23917/arstech.v4i2.2576

Article in Press

ARTICLE INFO

Article history:

Received 23 August 2023
Revised 08 September 2023
Accepted 15 September 2023
Available online 30 November 2023
Published regularly 31 December 2023

Keywords:

Anomaly Detection
Deep Learning
Exponential Smoothing
Long Short-Term Memory (LSTM)
Market Manipulation

ABSTRACT

Anomaly detection in time series data is a complex data mining issue with many useful, real-world applications. Anomalies in datasets represent deviations in the expected behaviour of a system and can indicate rare but significant events that require intervention. Market manipulation is a serious issue in financial jurisdictions worldwide, with financial regulators such as the SEC constantly trying to prevent it and prosecute those guilty of it. This paper makes use of state-of-the-art deep learning techniques as well as more classical statistical techniques in order to detect anomalies in five real-world datasets. The predictions of these models are then aggregated in two different ensemble models. The results of the individual models as well as the ensemble models, are evaluated, and F1-Score measures performance. Nine individual models, consisting of three models based on LSTM with Dynamic Thresholding, three ARIMA models and three Exponential Smoothing models, were used to generate predictions of anomalies based on daily trading volumes. The individual predictions of these models were then aggregated, with two different ensemble methods being used, namely the majority voting ensemble method and the ensemble averaging aggregation method. While both performed well, the majority voting ensemble method was seen to be the superior method in this study, with an average F1Score of 0.494, compared to an F1Score of 0.414 for the ensemble averaging aggregation method.

2722-9642 / © 2023 The Author(s). Published by Universitas Muhammadiyah Surakarta.
This is an open-access article under the CC BY license (https://creativecommons.org/licenses/by/4.0)

1. INTRODUCTION

Anomaly detection is a commonly practised task within the data mining and machine learning fields due to its applicability to many applications. An anomaly can be defined as "an observation that deviates so significantly from other observations as to arouse suspicion that it was generated by a different mechanism" [1]. Depending on the context, this observation could signal several rare occurrences, such as manufacturing defects [2], impending criminal activity [3] or gambling fraud [4]. In many fields, the early identification of these anomalies is vital as it allows for intervention which can mitigate the ill effects arising from these anomalous conditions.

Anomalies can generally be placed into one of three categories: Point anomalies, Collective anomalies and Contextual Anomalies. Point anomalies refer to data points considered abnormal when viewed against all other data points in a dataset. Collective anomalies are groups of data points considered abnormal when viewed against the whole dataset, even if the individual points are not abnormal. Contextual anomalies are data points considered abnormal when viewed against nearby or neighbouring data points. However, they may not be abnormal when viewed against the entire dataset [5].

One increasingly studied application of anomaly detection systems is identifying market manipulation or fraud within financial trading systems such as stock markets. A stock market is a place where participants can engage in the trading of stocks (equity) and other financial instruments of publicly listed companies [6]. Trade regulations exist in most countries, which are designed to maintain fairness between buyers and sellers and ensure the efficiency of the market. However, some bad actors will breach these regulations and aim to manipulate stock prices (increase or decrease) for personal gains. In 1934, the United States Congress formed the Securities and Exchange Commission (SEC) partly to combat and eliminate stock market manipulation. While stock market manipulation has decreased, it remains a severe issue in the Over-The-Counter (OTC) market within the United States and other financial jurisdictions [7]. The SEC (U.S. Securities and Exchange Commission) defines market manipulation as "Intentional or wilful conduct designed to deceive or defraud investors by controlling or artificially affecting the price of securities, or the intentional interference with the free forces of supply and demand" [8]. Due to the vast number of daily trades (~30,000,000 in NASDAQ), it is not feasible for humans to manually search for and detect anomalous patterns in stock market data, which could indicate market manipulation. For this reason, some attempts have been made to implement machine-learning approaches capable of quickly identifying and flagging anomalies in stock market data.

In recent years, advancements in machine learning have enabled researchers to apply deep learning approaches to financial data with the goal of identifying anomalies. The term' deep learning' refers to approaches based on artificial neural network architectures that have been applied to a wide variety of domains such as automated vehicles [9], language processing [10] and medical research [11]. This paper aims to use an ensemble method combining deep learning and more classical statistical approaches to identify anomalies in five real-world data sets which point to stock market manipulation. The approaches will be unsupervised, meaning anomalies will not be labelled to mirror real-world conditions most accurately. This study leverages a popular deep learning method known as Long Short-Term Memory (LSTM) with Dynamic Thresholding, as well as the statistical methods Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing to identify anomalies in the given data sets. The proposed framework then combines the predictions of the individual models with two different methods trialled to produce a better anomaly detection system.

2. RELATED WORKS

In recent years, anomaly detection approaches involving deep learning have enjoyed increased attention due to their high-performance levels across several domains. Chalapathy et al. [12] presented a structured and comprehensive overview of deep learning-based anomaly detection research methods. This review includes anomaly detection based on time series data and images for applications ranging from fraud detection to medical anomaly detection. Chandola et al. [13] provided a similar review of anomaly detection techniques, including reviews of both parametric and nonparametric statistical techniques, as well as simple machine learning models such as k-nearest neighbours. Whilst both studies are comprehensive in their techniques, no comparison is made between techniques to gauge relative performance. Goldstein and Uchida [14] provided an analysis and comparison of 19 different unsupervised anomaly detection algorithms tested on ten different multivariate datasets from multiple application domains. The statistical method Histogram-Based Outlier Score (HBOS), as well as K-Nearest Neighbour and Local Outlier Factor, were found to be the algorithms which performed best, with HBOS noted as being particularly useful for large datasets due to its low computational cost.

Some papers have also focused more specifically on anomaly detection of stock market data. Islam et al. [15] proposed the ANOMALOUS algorithm, which analysed historical stock market data detailing the daily volume traded of the stocks related to a group of companies. These companies were chosen as they were involved in

Article in Press

legal cases relating to insider trading (the practice of purchasing or selling a publicly traded company's securities while possessing material information that is not public) [16]. The ANOMALOUS algorithm utilised a Long Short-Term Memory Recurrent Neural Network (LSTM RNN) to predict stock transaction volume, then compared these predictions with the real-world data and used these comparisons to declare whether or not a point was anomalous. The algorithm showed success in identifying the periods where training volumes were deemed to be anomalous. Still, comparison to other state-of-the-art work was impossible as this was the first paper to detect illegal insider trading from real-world legal cases specifically.

LSTMs have become increasingly popular within all fields of anomaly detection. Leangarun et al. [17] also focused on detecting instances of stock market manipulation, developing an unsupervised approach using Long Short-Term Memory Generative Adversarial Networks (LSTM-GANs). LSTM was a base structure of GANs, which learned normal market behaviours. Once trained, the discriminator network of the GAN was used as a detector to differentiate between normal and anomalous trading patterns. Unlike the previously discussed study, this model was trained on normal data, with simulated cases of market manipulation used for testing. 68.1% accuracy was achieved on the unseen testing data, in which the focus was the detection of "pump-and-dump" (a manipulative scheme that attempts to boost the price of a stock or security through fake recommendations) schemes [18].

This paper will make use of a method proposed by Hundman et al. [19], which was designed to detect anomalies in spacecraft monitoring systems. LSTMs were used to predict the spacecraft's telemetry, after which a novel dynamic thresholding approach was used to identify anomalous points. This method was applied to stock market anomaly detection by Tallboys et al. [20]. A real-world dataset consisting of the daily trading volumes of five companies identified by the SEC as illegal was produced. Two deep-learning algorithms were used to detect these anomalies after being trained on two years of average trading data. The previously mentioned LSTM with Dynamic Thresholding TadGAN was also used. TadGAN is an unsupervised anomaly detection approach developed and presented by Geiger et al. [21] built on Generative Adversarial Network (GAN) architecture, with LSTM RNNs used as base generators and critic models. Tallboys et al. [20] used these two deep learning models to detect anomalies in the real-world dataset and compared the results with the classical statistical approach ARIMA. The results showed that ARIMA produced the best results as measured by F1 score in four out of five cases, with the LSTM with Dynamic Thresholding approach being the better of the two deep learning methods. This LSTM with a Dynamic Thresholding approach proved particularly useful where the anomalies present in a dataset are contextual/local.

Buda et al. [22] developed the IBM DeepAD framework, which combines the predictions of state-of-the-art deep learning models like LSTM with other probabilistic and statistical models such as ARIMA and Holt-Winters Exponential Smoothing. Each model is trained and makes predictions of future data points separately. After this, the model's predictions are combined, with two different methods of combination trialled, one where an aggregate of all predictions is used and one where the model with the best-perceived performance is used solely. Finally, a dynamic threshold determined based on the squared error is used to declare data points as anomalous or not. DeepAD was benchmarked against the EGADS framework developed by Laptev et al. [23] and was found to generally outperform EGADS as measured by early detection score, precision, recall and F1-score.

3. DATASET DESCRIPTION

The dataset used for the analysis presented in this paper is the one proposed by Tallboys et al. [20]. The dataset consists of daily trading volume relating to five companies whose stock is publicly traded within the US markets. Four companies were identified by the researchers in publicly available documents released by the SEC as having been suspected of being involved in market manipulation. These documents include exact dates of the anomalous trades as well as details such as dates of trade suspensions. The fifth company included is the US retailer GameStop.

In early 2021, GameStop stock was involved in a widely publicised "short squeeze" (a phenomenon in financial markets where a sharp rise in the price of an asset forces traders who previously sold short to close out their positions) [24] led by the Reddit community r/wallstreetbets. For each company in the data set, a period of anomalous trading behaviour was identified, either from the publicly available SEC documentation or, in the case of GameStop, decided by the researchers. Data from the 24 months before the identified market manipulation was included, as was data from the 12 months after the anomalous period. The described data was publicly available and downloaded from the Yahoo! Finance API using the finance Python library. Table 1 shows details of the identified periods of market manipulation for each stock.

Figures 1-5 show the daily trading volumes for the five stocks in the dataset, with the anomalous periods highlighted.

Article in Press

Table 1. Companies included in the dataset and anomaly details

Company name	Stock ticker	Anomaly start	Anomaly end	Anomaly type
Aethlon Medical, Inc.	aemd	22-Jan-20	07-Feb-20	Contextual
Applied Biosciences Corp.	appb	25-Mar-20	13-Apr-20	Point
GameStop	gme	11-Jan-21	29-Jan-21	Point
No Borders, Inc.	nbdr	11-Mar-20	03-Apr-20	Point
Turbo Global Partners, Inc.	trbo	30-Mar-20	09-Apr-20	Point

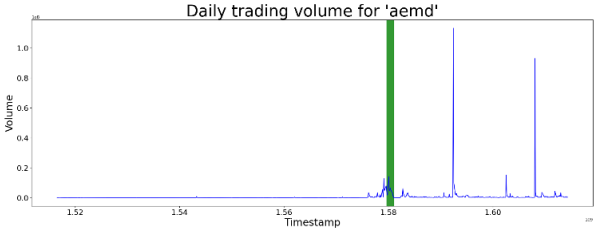

Figure 1. Daily trading volume for 'aemd'.

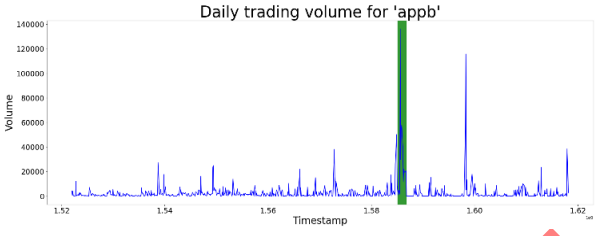

Figure 2. Daily trading volume for 'appb'.

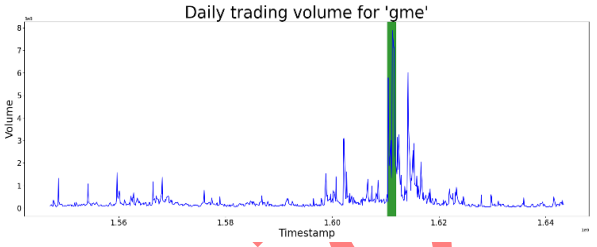

Figure 3. Daily trading volume for 'gme'.

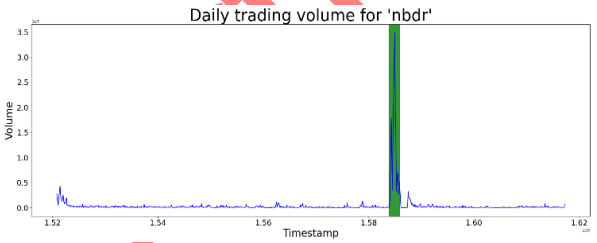

Figure 4. Daily trading volume for 'nbdr'.

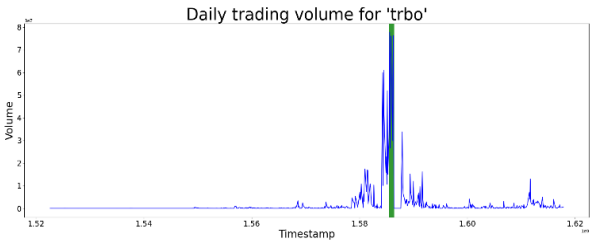

Figure 5. Daily trading volume for 'trbo'.

This paper is novel in that it is the first to use an unsupervised approach by using Long Short-Term Memory Generative Adversarial Networks (LSTM GANs) and applying it to daily trading volume whose stock is publicly traded.

4. METHODS

4.1. Overview of Developed Methodology

A visual representation of the developed methodology of this analysis is depicted in Figure 6 and similar to the approach taken by Buda et al. [22] when developing the DeepAD framework; this methodology consists of three main phases. The first phase involves the generation of individual anomaly predictions by nine different models (three LSTM models, three ARIMA models and three Exponential Smoothing models). The second phase will concern the aggregation of the predictions of these models with two different types of aggregation trialled. The final phase of the analysis will involve the ultimate detection of anomalies based on the previously merged individual predictions. The relative performance of the two merging systems, as well as the performance of the individual models, will then be compared and contrasted.

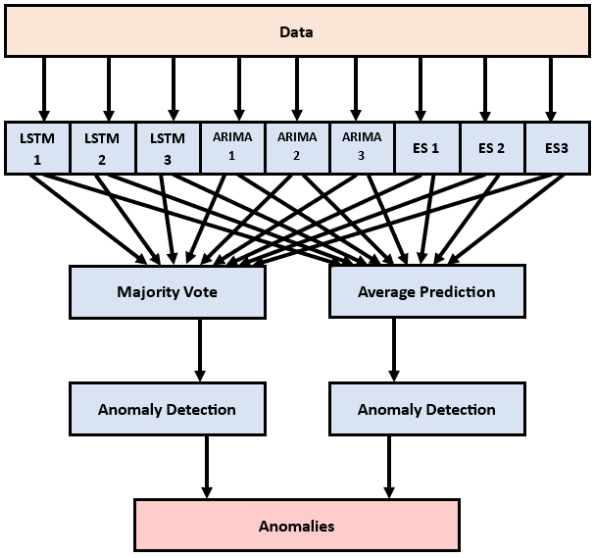

Figure 6. Overview of methodology.

In Press

4.2. Individual Models

LSTM with Dynamic Thresholding

This previously discussed approach was proposed by Hundman et al. [19] and was proven to be useful in the detection of real-world stock market manipulation by Tallboys et al. [19]. This approach uses an LSTM to generate predictions for future data points, followed by a nonparametric, unsupervised approach to find an error threshold to identify anomalies. LSTMs are a special kind of Recurrent Neural Network (RNN) that are capable of considering both recent and more long-term dependencies in data through the use of feedback connections. These feedback connections allow LSTMs to consider entire sequences of data, where each point is considered not independently but rather as part of the series of points that came before it, thus allowing the LSTM to learn sequential patterns in the data. This ability makes LSTMs incredibly useful when predicting future sequences in highly complex temporal data. The threshold used when determining a point to be anomalous or not considers the smoothed errors from the LSTM predictions and includes a pruning step that ensures only the anomalies that cause the greatest fluctuations in the mean and standard deviation of the smoothed errors are flagged. The code used to implement this approach was developed by the open-source machine-learning Python library Orion [25]. Three LSTM models were leveraged in this analysis, with the window size parameter being varied for each one. This parameter controls the length of the input sequence that the LSTM uses to predict future data points. LSTMs with window sizes of 250, 100 and 10 were trialled. The number of epochs was kept at a constant value of 10 to keep the models computationally cheap without much loss of accuracy, as models generally converged around this point or soon after. The default batch size of 64 was used in all three models.

ARIMA

ARIMA is a common time series forecasting model that captures the linear dependency of the future values on the previous data [26]. The working of an ARIMA model can be understood by breaking down its acronym. "AR" stands for auto-regressive, which means that the current or future values are correlated with values at previous time steps. "I" stands for integrated, which allows the model to handle non-stationary data by predicting the differences of the series from one time step to the next time step instead of the time series values themselves. "MA" stands for moving average, which analyses the error levels of previous predictions to make better predictions for future points. Tallboys et al. [20] used ARIMA as a benchmark for performance comparison with the two deep learning methods trialled and found that ARIMA was only outperformed in the case of one stock out of five. Once again, the code used to implement this approach was developed by the open-source machine-learning Python library Orion [25].

Similar to the previously described LSTM model, three ARIMA models were implemented with varying window sizes. The trialled window sizes were set at 100, 250 and 400.

Exponential Smoothing

Exponential smoothing is one of the most popular methods for smoothing discrete time series in order to forecast the immediate future [27]. This is likely due to its simplicity, low computational cost as well as its good general accuracy. Exponential smoothing works on the assumption that the most recent observations in a time series data set have a more significant influence on the forecast of future values than the more distant observations. Exponential Smoothing is a simple and pragmatic approach to forecasting whereby the forecast is constructed from an exponentially weighted average of past observations. The most enormous weight is given to the present observation, with less weight being given to the immediately preceding observation and even less weight to the observation before that and so on. This process leads to the exponential decay of the influence of a data point as it becomes more and more distant in the past [27–28]. A parameter known as the smoothing factor is responsible for determining how quickly the influence of a past point decays. As the smoothing factor increases (to a maximum of 1), the forecast becomes more reactive as points further in the past have a reduced influence. In contrast, a low smoothing factor (minimum of 0) is more highly influenced by distant points, which leads to a smoother, less reactive forecast [29].

As with the other two models previously described, three Exponential Smoothing models were implemented, with the smoothing factor being varied between the three. Smoothing factors of 0.025, 0.05 and 0.1 were used to generate the model predictions.

For the Exponential Smoothing model to decide if a point was anomalous, an upper and lower bound was created that would fluctuate in line with the model's predictions. If a real-world data point were above the upper bound or below the lower bound, that point would be classed as anomalous. In order to calculate the distance of the upper and lower bounds from the model prediction, Chebyshev's Inequality was leveraged. As the distribution of the real-world data used in this analysis is unknown, it was impossible to use the classic 68-95-99 rule (also known as the empirical rule), which only applies to normally distributed data. Chebyshev's Inequality, however, dictates that for a broad class of probability distributions, no more than a certain fraction of values can be more than a certain distance from the mean [22]. Kaban [30] showed that for a finite sample of N=500,

Article in Press

95% of all points should lie within 4.5774 standard deviations. Any point further than 4.5774 standard deviations (plus the mean absolute error of the predictions) was classed as anomalous.

4.3. Aggregation of Models

In a similar fashion to Buda et al. [22], the second phase of the proposed method involves combining the outputs of the nine previously mentioned individual models. Two different methods of aggregation will be trialled.

Majority Vote

This method is a majority voting ensemble and will take a single vote from each of the nine individual models for each data point in the time series as a Boolean variable. The Boolean variable will denote the perceived status of that point with '1', meaning that the point is considered anomalous and '0', meaning the point is not considered anomalous. The final prediction is then given by taking the majority vote, i.e., the class (anomalous or not anomalous) that obtains the highest number of votes (the most frequent vote) is passed as the final prediction [31].

Average Prediction

This method aims to implement ensemble averaging by finding the mean average of the predicted values by each individual for each data point in the time series. For each timestep in the time series, the individual models will produce a predicted value based on that individual model type and the parameters used. The final prediction of this aggregation method will be the mean average of these individual model predictions.

4.3. Identification of Anomalies

The final stage of the process involves the detection of perceived anomalies in the dataset. The anomaly detection process following each aggregation method will differ due to the different types of outputs.

Majority Vote

In the case of the majority voting ensemble model, a point will be declared anomalous or not based on the consensus of the individual models. If a majority (five or more) of the models return an anomalous verdict for a given data point in the time series, then that point will be declared anomalous.

Average Prediction

The procedure for deciding if a given data point in the time series is anomalous or not in the case of the ensemble averaging model will be similar to the procedure discussed earlier when describing the Exponential Smoothing model. The 68-95-99 rule cannot be used because the data distribution is unknown, so Chebyshev's Inequality is used. Any data point in the time series that is more than 4.5774 standard deviations (plus the mean absolute error of the predictions) away from the average prediction value will be declared anomalous.

5. RESULTS AND DISCUSSION

As in previous studies related to anomaly detection [19–22], the basis for the calculation of evaluation metrics is as follows:

- True Positive (TP) if a data point is predicted to be anomalous and also falls within the real-world identified anomaly window.
- False Positive (FP) if a data point is predicted to be anomalous and does not fall within the real-world identified anomaly window.
- True Negative (TN) if a data point is predicted not to be anomalous and does not fall within the real-world identified anomaly window.
- False Negative (FN) if a data point is predicted not to be anomalous but does fall within the real-world identified anomaly window.

These metrics will be used to calculate the principal evaluation metrics for this analysis, those being Precision, Recall and F1-Score. Precision is the ratio of true positives (TP) to total positives, or in other words, how often a model is correct when it makes a positive prediction. Recall measures how well a model predicts true positives, or in other words, what percentage of real-world positives were correctly predicted by the model. The Orion library computes the F1Score using the weighted segment approach proposed by Alnegheimish et al. [32], which works by first splitting the time series into multiple sequences by the edges of the anomalous intervals. After this, a confusion matrix is constructed, which makes a segment-to-segment comparison and records true positive, false positive, false negative, and true negative accordingly, then weights each segment by its duration [33]. This approach penalises the predictive model when the detected window differs in any way from the real anomalous window as opposed to an overlapping segment approach, which grants a true positive if a real anomaly falls anywhere in the detected window [20].

This section of the analysis will be broken down into two subsections. Firstly, the performance of the individual models on each of the five stocks involved in the dataset will be evaluated and compared. F1-Score will be used as the sole evaluation metric for the individual models. After this, the success of each of the aggregation methods will be evaluated and compared with both each other's and the individual models'. Although F1-Score will be the principal evaluation metric for the aggregation models, Precision and Recall will also be used to explain the differing performances of the two methods.

Article in Press

5.1. Individual Models

Table 2 shows the performance of each individual model, as measured by F1-Score when attempting to detect anomalies in the five real-world datasets, as well as the model's average score across all tests. The model which performed best on the data for each stock is also highlighted.

As was the case with the findings of Tallboys et al. [20] the ARIMA models are seen to perform the best, with an ARIMA-based model returning the highest F1-score for four of the five stocks. When an average of the scores for each model across the five stocks is taken, the three highest-scoring models in terms of F1-Score are the three ARIMA models implemented. The ARIMA models with the window size parameter set to 100 and 400 performed equally well, tying for the best overall individual model. However, all of the ARIMA models failed to identify the contextual anomaly present in the 'aemd' dataset.

For the 'aemd' dataset, the only group of models that detected the real-world contextual anomalous period with any level of success were the three that leveraged LSTM with Dynamic Thresholding. This is particularly promising as this anomaly was not particularly large in terms of volume when compared with other legitimate fluctuations that coincided with FDA approvals [34] and clinical trials [35] related to the company (Aethlon Medical, Inc.). The flexibility of the LSTM models is evidenced by the fact that the LSTM-based models with window sizes 250 and 10 were the only models which were able to detect the real-world contextual anomalous period with any level of success across all five stocks included in this analysis. In terms of overall average performance, the LSTM with Dynamic Thresholding model with window size parameter set at 10 performed the best of the three, but only marginally.

All three of the Exponential Smoothing models, like ARIMA, failed to identify the contextual anomalous period within the 'aemd' data. This is unsurprising because, as previously discussed, the anomalous points were at relatively low volumes compared to other points in the dataset. The fact that the anomaly detection limit for this model type is based on a multiple (4.5774) of the standard deviation of the predicted values means that any large fluctuations in trading volume will serve to push the boundaries further away and make detection of anomalies of smaller amplitudes more difficult. When dealing with the other four datasets, the Exponential Smoothing models performed well but fell short of the performance exhibited by the ARIMA models in all cases.

Figures 7-11 show the daily trading volume of each of the five stocks overlayed with the predictions of the best-performing individual model (real anomaly highlighted in green, predicted anomaly highlighted in red).

Table 2. F1-Scores for each individual model and dataset.

Model	Stock Name					
	aemd	appb	gme	nbdr	trbo	average
lstm_250	0.17	0.39	0.58	0.36	0.06	0.31
lstm_100	**0.29**	0.32	0.26	0.69	0.00	0.31
lstm_10	0.24	0.36	0.31	0.50	0.22	0.33
arima_100	0.00	**0.51**	**0.65**	**0.93**	0.63	**0.54**
arima_250	0.00	0.47	0.61	0.87	0.71	0.53
arima_400	0.00	0.48	0.62	0.85	**0.77**	**0.54**
ES_0.025	0.00	0.36	0.48	0.51	0.56	0.38
ES_0.05	0.00	0.30	0.48	0.47	0.56	0.36
ES_0.1	0.00	0.08	0.47	0.42	0.67	0.33

Article in Press

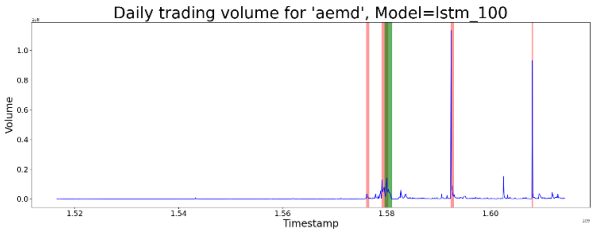

Figure 7. LSTM with Dynamic Thresholding (window size = 100) prediction for 'aemd' dataset.

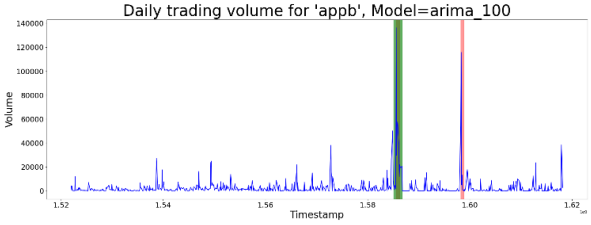

Figure 8. ARIMA (window size = 100) prediction for 'appb' dataset.

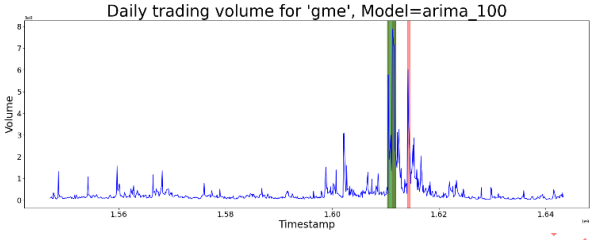

Figure 9. ARIMA (window size = 100) prediction for 'gme' dataset.

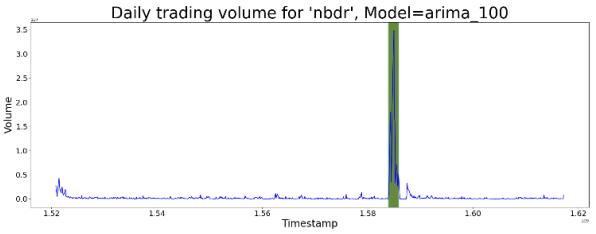

Figure 10. ARIMA (window size = 100) prediction for 'nbdr' dataset.

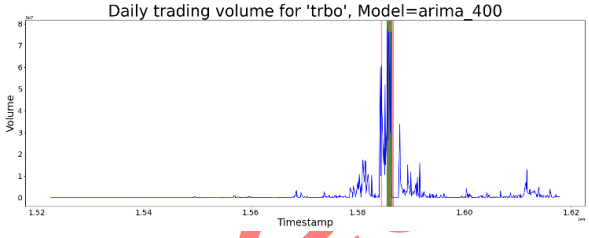

Figure 11. ARIMA (window size = 400) prediction for 'trbo' dataset.

5.2. Aggregation of Models

Table 3 shows the performance results of the two aggregation models used and includes details of the Precision, Recall and F1-Scores.

When comparing the performance of the two model aggregation methods, it is immediately apparent that the majority voting ensemble method has resulted in better F1 scores in almost all cases. A higher average F1-Score was achieved in four out of five cases, with the 'aemd' dataset being the exception, as both models scored zero. Whilst this level of performance is encouraging, it is noted that the majority vote aggregation method only outperformed the best individual model in one out of the five tests, when the 'tbro' dataset was used (F1-Score=0.80).

Table 3. Precision, Recall and F1-Scores for each aggregation model & dataset.

Aggregation model	Measure	Stock Name					
		aemd	appb	gme	nbdr	trbo	Average
Vote	Precision	0	0.53	0.63	1	0.75	0.582
	Recall	0	0.4	0.48	0.5	0.86	0.448
	F1-Score	0	**0.46**	**0.54**	**0.67**	**0.8**	**0.494**
Average	Precision	0	0.63	0.75	1	0.79	0.634
	Recall	0	0.25	0.29	0.23	0.79	0.312
	F1-Score	0	0.36	0.41	0.51	0.79	0.414

Article in Pre

While the ensemble averaging aggregation method also performed fairly well, it failed to beat the majority voting ensemble method for any dataset in terms of F1-Score. The reason for this can be seen in Table 3. The ensemble averaging was actually found to have outperformed the majority voting ensemble method in terms of precision but fell far short when recall measurements were compared. This means that although the model predicts anomalies with a high degree of confidence, it fails to do it often enough, which results in a comparatively high number of false negatives. One potential way to improve the recall value could be to reduce the number of standard deviations away from the average predicted value that a data point must be to be considered anomalous. However, this could have negative impacts on the precision score as more false positives are likely to occur as well as the intended true positives.

The fact that neither of the two model aggregation methods could correctly identify the anomalous window in the 'aemd' dataset represents a significant limitation of this study. That this is the case is unsurprising, as six of the nine individual models that contributed to the aggregation models could not identify the anomalous window. This is an area that could potentially be improved upon in further work, with one possibly viable solution being that, in some instances, the number of votes required to produce a positive (anomalous) prediction from the voting ensemble method could be reduced so as an absolute majority is not required, instead using a lower threshold, e.g., 33%.

Overall, these results should be viewed positively, as they have exceeded the benchmarks set in previous literature [20], both in terms of the individual and ensemble models. Models such as these could be reasonably seen to have real-world applications as warning systems for financial regulators such as the SEC. Models such as these could be run daily or weekly and produce a list of companies potentially guilty of market manipulation. A human analyst could then analyse the stock in more detail to identify whether there are legitimate reasons for the anomalous points.

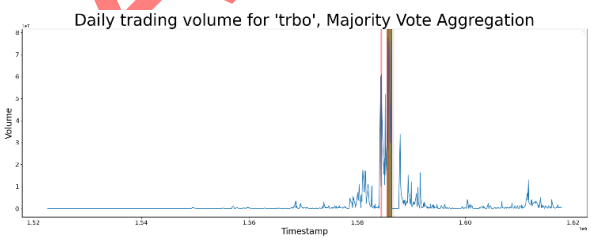

Figure 12. Majority Vote Aggregation Model prediction for 'trbo' dataset.

Figure 12 shows the daily trading volume of the 'trbo' stock overlayed with the predictions of the majority voting aggregation model, as it was the only aggregation model to outperform all of the individual models on a particular stock (real anomaly highlighted in green, predicted anomaly highlighted in red).

4. CONCLUSION

This paper used the five real-world, labelled datasets proposed by Tallboys et al. [20] and applied a mixture of deep learning and more classical, statistical techniques to detect anomalies in these datasets. Nine individual models, consisting of three models based on LSTM with Dynamic Thresholding, three ARIMA models and three Exponential Smoothing models, were used to generate predictions of anomalies based on daily trading volumes. The individual predictions of these models were then aggregated, with two different ensemble methods being used, namely the majority voting ensemble method and the ensemble averaging aggregation method. While both performed well, the majority voting ensemble method was considered the superior method in this study, with an average F1-Score of 0.494, compared to an F1-Score of 0.414 for the ensemble averaging aggregation method. This proves the real-world viability as well as the flexibility of these models. Both ensemble models and the individual models developed surpassed the current benchmarks in literature for this dataset. Unfortunately, six individuals and both aggregation models could not detect the contextual anomaly in one of the datasets. This issue should form the basis for future work, with methods developed to identify contextual anomalies more readily, especially those that do not deviate as drastically from the mean prediction as other points in the same dataset.

CONFLICTS OF INTEREST

The author declares that no competing financial interests could have appeared to impact the work.

REFERENCES

[1] D.M. Hawkins, "Identification of outliers", *Monographs on Statistics and Applied Probability*, vol. 11, 1980. https://doi.org/10.1007/978-94-015-3994-4

[2] R.J. Hsieh, J. Chou, and C.H. Ho, "Unsupervised online anomaly detection on multivariate sensing time series data for smart manufacturing", *Proceedings - 2019 IEEE 12th Conference on Service-Oriented Computing and Applications, SOCA 2019*, Institute of Electrical and Electronics Engineers Inc., pp. 90–97, 2019. https://doi.org/10.1109/SOCA.2019.00021

Article in Press

[3] S. Chackravarthy, S. Schmitt, and L. Yang, "Intelligent crime anomaly detection in smart cities using deep learning", *Proceedings - 4th IEEE International Conference on Collaboration and Internet Computing, CIC 2018*, Institute of Electrical and Electronics Engineers Inc., pp. 399–404. 2018. https://doi.org/10.1109/CIC.2018.00060

[4] M. Min, J.J. Lee, H. Park, H. Shin, and K. Lee, "A statistical approach towards fraud detection in the horse racing", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, pp. 191–202, 2020. https://doi.org/10.1007/978-3-030-65299-9_15

[5] M.A. Hayes and M.A. Capretz, "Contextual anomaly detection framework for big sensor data", *Journal of Big Data*, vol. 2, no. 1, 2015, https://doi.org/10.1186/s40537-014-0011-y

[6] I.K. Nti, A.F. Adekoya, and B.A. Weyori, "A systematic review of fundamental and technical analysis of stock market predictions", *Artificial Intelligence Review*, vol. 53, no. 4, pp. 3007–3057, 2020. https://doi.org/10.1007/s10462-019-09754-z

[7] R.K. Aggarwal and G. Wu, "Stock market manipulations", *Journal of Business*, vol. 79, no. 4, pp. 1915–1953, 2006. https://doi.org/10.1086/503652

[8] US Securities and Exchange Commission, "Market manipulation and case studies", 2023. https://www.sec.gov/file/market-manipulation-and-case-studies

[9] B. Sairam, A. Agrawal, G. Krishna, and S.P. Sahu, "Automated vehicle parking slot detection system using deep learning", *Proceedings of the 4th International Conference on Computing Methodologies and Communication, ICCMC 2020*, Institute of Electrical and Electronics Engineers Inc., pp. 750–755, 2020. https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000140.

[10] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing", *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018. https://doi.org/10.1109/MCI.2018.2840738

[11] X. Chen, X. Wang, K. Zhang, K.M. Fung, T.C. Thai, K. Moore, R.S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis", *Medical Image Analysis*, vol. 79, p. 102444, 2022. https://doi.org/10.1016/j.media.2022.102444

[12] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey", *Computer Science*, 2019. http://arxiv.org/abs/1901.03407

[13] N.R. Prasad, S. Almanza-Garcia, and T.T. Lu, "Anomaly detection", *Computers, Materials and Continua*, vol. 14, no. 1, pp. 1–22, 2009. https://doi.org/10.1145/1541880.1541882

[14] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data", *PLoS One*, vol. 11, no. 4, 2016. https://doi.org/10.1371/journal.pone.0152173

[15] S.R. Islam, S. Khaled Ghafoor, and W. Eberle, "Mining illegal insider trading of stocks: a proactive approach", *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, Institute of Electrical and Electronics Engineers Inc., pp. 1397–1406. 2019. https://doi.org/10.1109/BigData.2018.8622303

[16] CFI for Team, "Insider Trading", 2023. https://corporatefinanceinstitute.com/resources/wealth-management/what-is-insider-trading/

[17] T. Leangarun, P. Tangamchit, and S. Thajchayapong, 'Stock price manipulation detection using generative adversarial networks', *IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 2104–2111, 2018. https://doi.org/10.1109/SSCI.2018.8628777

[18] R. Dhir, "Pump-and-dump: Definition, how the scheme is illegal, and types", *Investopedia*, 2022. https://www.investopedia.com/terms/p/pumpanddump.asp

[19] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding", *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Association for Computing Machinery*, pp. 387–395. 2018. https://doi.org/10.1145/3219819.3219845

Article in Press

[20] J. Tallboys, Y. Zhu, and S. Rajasegarar, "Identification of stock market manipulation with deep learning", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Science and Business Media Deutschland GmbH, pp. 408–420, 2022. https://doi.org/10.1007/978-3-030-95405-5_29.

[21] A. Geiger, D. Liu, S. Alnegheimish, A. Cuesta-Infante, and K. Veeramachaneni, "TadGAN: Time series anomaly detection using generative adversarial networks", *Computer Science*, 2020. http://arxiv.org/abs/2009.07769

[22] T.S. Buda, B. Caglayan, and H. Assem, "DeepAD: A generic framework based on deep learning for time series anomaly detection", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2018, pp. 577–588. https://doi.org/10.1007/978-3-319-93034-3_46.

[23] N. Laptev, S. Amizadeh, and I. Flint, "Generic and scalable framework for automated time-series anomaly detection", *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, pp. 1939–1947, 2015. https://doi.org/10.1145/2783258.2788611.

[24] CFI for Team, "Short squeeze", 2023. https://corporatefinanceinstitute.com/resources/career-map/sell-side/capital-markets/short-squeeze/

[25] GitHub, "Sintel-dev/Orion", 2023. https://github.com/sintel-dev/Orion

[26] E.H.M. Pena, M.V.O. De Assis, and M.L. Proença, "Anomaly detection using forecasting methods ARIMA and HWDS", *Proceedings - International Conference of the Chilean Computer Science Society, SCCC*, IEEE Computer Society, pp. 63–66, 2013. https://doi.org/10.1109/SCCC.2013.18

[27] E. Ostertagova, O. Ostertag, and E. Ostertagová, "The simple exponential smoothing model", *The 4th International Conference on modelling of mechanical and mechatronic systems, Technical University of Košice, Slovak Republic, Proceedings of Conference*. p. 380-384, 2011.

[28] A.D. Aczel, "Complete business statistics", McGraw Hill, 1998.

[29] D.C. Montgomery, L.A. Johnson, and J.S. Gardiner, "Forecasting and time series analysis", McGraw-Hill, Inc., 1990.

[30] A. Kabán, "Nonparametric detection of meaningless distances in high dimensional data", *Statistics and Computing*, vol. 22, no. 2, pp. 375–385, 2012. https://doi.org/10.1007/s11222-011-9229-0

[31] A. Dogan and D. Birant, "A weighted majority voting ensemble approach for classification", *Proceedings, 4th International Conference on Computer Science and Engineering*, Institute of Electrical and Electronics Engineers Inc., pp. 366–371, 2019. https://doi.org/10.1109/UBMK.2019.8907028

[32] S. Alnegheimish, D. Liu, C. Sala, L. Berti-Equille, and K. Veeramachaneni, "Sintel: A machine learning framework to extract insights from signals", in Proceedings of the ACM SIGMOD International Conference on Management of Data, Association for Computing Machinery, pp. 1855–1865, 2022. https://doi.org/10.1145/3514221.3517910

[33] S. Alnegheimish, 'Orion-a machine learning framework for unsupervised time series anomaly detection', PhD Thesis. Massachusetts Institute of Technology, 2022. https://dai.lids.mit.edu/wp-content/uploads/2022/06/sarah_sm_thesis.pdf

[34] Aethlon Medical, "Aethlon announces FDA approval of IDE supplement for COVID-19 patients", *CISION PR Newswire*, 2020. https://www.prnewswire.com/news-releases/aethlon-announces-fda-approval-of-ide-supplement--for-covid-19-patients-301079557.html

[35] Aethlon Medical, "Aethlon medical announces first patient treated in first-in-human clinical trial of HEMOPURIFIER® in head and neck cancer", *CISION PR Newswire*, 2020. https://www.prnewswire.com/news-releases/aethlon-medical-announces-first-patient-treated-in-first-in-human-clinical-trial-of-hemopurifier-in-head-and-neck-cancer-301193962.html

Article in Press