Detection of Stock Market Manipulation Using Deep Learning

Adarsh Kumar Singh[2], Jaideep Singh Garlyal[2], Karnik Kanojia[1], Debasmita Paul[1], Aadya Goel[1]

School of Computing Science and Engineering
[1] VIT Vellore University, Tamil Nadu, India
[2] SRM Chennai University, Tamil Nadu, India

***Abstract*- The stock market is a vast trading environment that handles millions of transactions, making it challenging for regulatory bodies to manually identify fraudulent activities. However, unsupervised deep learning techniques provide a promising solution for detecting market manipulation. Market manipulation occurs when traders manipulate stock prices to their advantage, either by inflating or deflating them. The paper proposes to investigate how market structure analysis can be used to detect stock market manipulation. Data was collected based on information provided by National Stock Exchange (NSE) and Bombay Stock Exchange (BSE) websites. The unsupervised generative models-based approach seems particularly promising in this domain as it can quickly identify contextual local anomalies in the data to detect anomalies and potential market manipulation, which can often be a challenge for deep learning approaches. As we address this issue, we hope to provide valuable insights that can help investors and regulators mitigate the risks associated with stock market manipulation.**

Keywords – Stock Market, Market Manipulation, Deep Learning

I. INTRODUCTION

Market manipulation poses a significant challenge to the integrity and fairness of financial markets. It involves the deliberate distortion of market prices, volumes, or other trading indicators to create artificial market conditions for personal gain. Detecting and preventing market manipulation is crucial for maintaining a transparent and trustworthy trading environment.

Traditional approaches for market manipulation detection rely on rule-based systems and statistical methods, which often struggle to capture the complex patterns and subtle manipulative behaviours prevalent in modern financial markets. However, recent advancements in deep learning have shown promise in addressing these challenges by leveraging the power of artificial neural networks to learn intricate patterns and relationships within data.

This paper focuses on the application of deep learning methods for market manipulation detection. Deep learning models, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Generative Adversarial Networks (GANs), have demonstrated remarkable success in various domains, including computer vision, natural language processing, and speech recognition. The ability of these models to automatically extract and learn complex features from raw data makes them promising candidates for enhancing market manipulation detection.

The objective of this research is to develop and evaluate deep learning-based models for effectively identifying and mitigating market manipulation. We aim to leverage the temporal and spatial dependencies present in stock market data to capture the nuanced patterns associated with manipulative activities. By utilising large-scale datasets containing real-world trading data, we can train and evaluate deep learning models on diverse manipulation scenarios, such as pump and dump, spoofing, and layering.

The contributions of this research lie in several aspects. Firstly, we comprehensively investigate the effectiveness of unsupervised deep learning methods for market manipulation detection. We explore generative deep learning architectures and techniques to identify the most suitable models for capturing manipulative behaviours. Secondly, we construct extensive datasets containing diverse manipulation scenarios, enabling robust training and evaluation of the proposed models. Finally, we evaluate the performance of the deep learning models using rigorous metrics, considering factors such as precision, recall, f1 score and generalisation capabilities across different market conditions.

The rest of the paper is organised as follows: Section 2 provides a review of related work on market manipulation detection and the application of deep learning in financial domains. Section 3 describes the methodology and the proposed deep learning models for market manipulation detection and also presents the experimental setup, including the datasets used and the evaluation metrics. Section 4 discusses the results and performance analysis of the deep learning models. Finally, Section 5 concludes the paper, by highlighting the contributions, limitations, and future research directions in the field of market manipulation detection using deep learning methods.

II. RELATED WORKS

Shashank Sridhar et. al [1] propose the utilisation of ensemble neural networks for detecting market manipulation. The authors' system can identify three different types of manipulation scenarios: price manipulation, volume manipulation, and trade reversal. To create the dataset, the authors gathered information from the Securities and Exchange Board of India (SEBI) and the Bombay Stock Exchange (BSE) website regarding 16 companies that had experienced stock manipulation. They combined the daily data from the pre-investigation period, investigation period, and post-investigation period, associating each chosen stock with a specific manipulation type. The dataset was then used to train and test the model. The researchers implemented an ensemble neural network model, both with and without trainable sub-model layers, on the daily trading dataset. The stacked model without trainable layers achieved the highest accuracy among the models. Overall, both models outperformed other supervised learning models in the evaluation.

T. Leangarun et al. [2] developed a Neural network to predict market manipulation. The system focused on two specific scenarios: data manipulation through pump and dump strategies, as well as spoof trading. Pump and dump involve artificially inflating a stock's price by purchasing it, while spoof trading involves deceiving other buyers into purchasing a specific stock at a particular price. The authors collected tick data from NASDAQ for three companies: Amazon, Intel, and Microsoft. They constructed a dataset using the tick trading data from these companies, specifically selecting Level 2 data because it contains valuable information about order cancellations, which is crucial for identifying pump and dump activities. The authors achieved a commendable 88.2% accuracy in detecting pump and dump manipulation through their developed Neural network. However, the network did not effectively model spoof trading. The Neural Network architecture consisted of 25 nodes in the input layer, three nodes in the hidden layer, and one node in the output layer.

K. Golmohammdi et. al [3] conducted a comparative analysis to predict market manipulation by employing various supervised learning algorithms. The researchers utilised a readily available dataset containing instances of market manipulation cases that occurred between January and December 2003. The Securities and Exchange Commission (SEC) served as the regulatory body responsible for determining whether a stock was manipulated or not. The dataset comprised 175,738 data observations from 64 issuers, encompassing 69 data attributes that represented parameters utilised in analytical analysis. The identified types of market manipulation included marking the close, wash trades, and cornering the market. Several supervised learning algorithms were employed to detect market manipulation, including CNN, Random Forest, C5.0, CTree, Neural Networks, CART, and Naive Bayes. In the conducted experiments, Naive Bayes outperformed all other supervised algorithms in terms of sensitivity and specificity when it came to identifying market manipulation.

A. Li et al. [4] utilises supervised learning algorithms to identify stock market manipulation in the Chinese market. The dataset comprises information from the China Securities Regulation Commission (CSRC) and security market data. The CSRC identified 64 manipulated stocks from 2013 to 2016, and the daily trading and tick trading data of these stocks were used. Classification methods were employed, including K Nearest Neighbour, Decision Tree, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Logistic Regression, Artificial Neural Networks, and Support Vector Machine. The experiments showed that these methods effectively detect market manipulation from daily trading data, while KNN and Decision trees performed the best among the algorithms used.

T. Leangarun et al. [5] implement Generative Adversarial Networks (GANs) to detect abnormal trading behaviours caused by stock price manipulations. Unlike other systems, it uses an unsupervised GAN with LSTM to learn abnormal market behaviour. The dataset consists of major companies from the Stock Exchange of Thailand (SET) over 22 trading days. The proposed hybrid model, combining GAN and LSTM, effectively detects anomalies, particularly in pump and dump manipulation cases. The system achieves a 68.1% accuracy in detecting market manipulation.

Q.Wang et al. [6] introduce a framework, called RNN-EL (RNN-based ensemble learning), for detecting stock market manipulation. The dataset used in the study consists of manipulation cases reported by the China Securities Regulation Commission (CSRC) from 2012 to 2016, addressing the challenges related to detecting trade-based stock price manipulation in China. The dataset includes 40 CSRC reports, involving 33 individual manipulators or groups, 64 stocks, and 257 manipulated cases. In contrast to other systems, the proposed framework incorporates trade-based features and characteristic facts of stocks to construct an RNN-EL model. A comparison was made with various supervised algorithms, demonstrating that the proposed model outperforms other methods in detecting market manipulation. These findings indicate the effectiveness of RNN-EL as a reliable mechanism for manipulation detection and highlight its ability to enhance detection capabilities.

J. Tallboys et al. [7] propose five large real-world, labelled data sets of anomalous stock market data where market manipulation is alleged to have occurred. It demonstrates that market manipulation can be detected using cutting-edge deep learning approaches, TadGAN and LSTM with Dynamic Thresholding, and with results compared with a more ARIMA approach. These were tested on stock market data with natural market manipulation anomalies. The data used was retrieved from the finance library in Python for 24 months before the identified market manipulation was selected. The algorithms were assessed on their efficiency, speed, breadth of anomaly types, and accuracy. The LSTM with Dynamic Thresholding was the most promising as it was able to detect local anomalies in data the fastest

III. METHODOLOGY

A. Dataset Description & Real Anomaly

Stock market data of companies was used to train models. Each company's history of manipulation was searched via search engines and articles from reputed media houses. Stock regulators like SEBI (Securities and Exchange Board of India) publicly share details of these cases and were checked for the period under which these stocks were investigated. Those periods were marked as anomalies. It is to note that data was limited to Indian markets only under the scope of research but the methodologies could easily be extended to any type of stock data.

Our original objective is to identify contextual anomalies. These observations are only anomalous relative to neighbouring data points but not an anomaly relative to all other observations and need sound reasons to be declared anomalous.

The data was downloaded from the official site of BSE (Bombay Stock Exchange) by searching for their Security ID/Name.

Sadhna Broadcast Ltd. The Securities and Exchange Board of India (SEBI) has found that the stock prices of Sadhna Broadcast Ltd were manipulated through misleading videos on some YouTube channels [11]. The videos falsely claimed that the company was going to be acquired by the Adani Group and had signed big contracts with Sony Pictures and Zee. This caused retail investors to buy the stock, driving up the price.

Sharpline Broadcast Ltd. Sharpline's stock was involved in the same scam along with Sadhna Broadcast Ltd [11].

The periods in which these stocks were under review by SEBI are provided in Table 1.

Name	Start	End
Sadhna Broadcast Ltd.	April 2022	September 2022
Sharpline Broadcast Ltd.	April 2022	August 2022

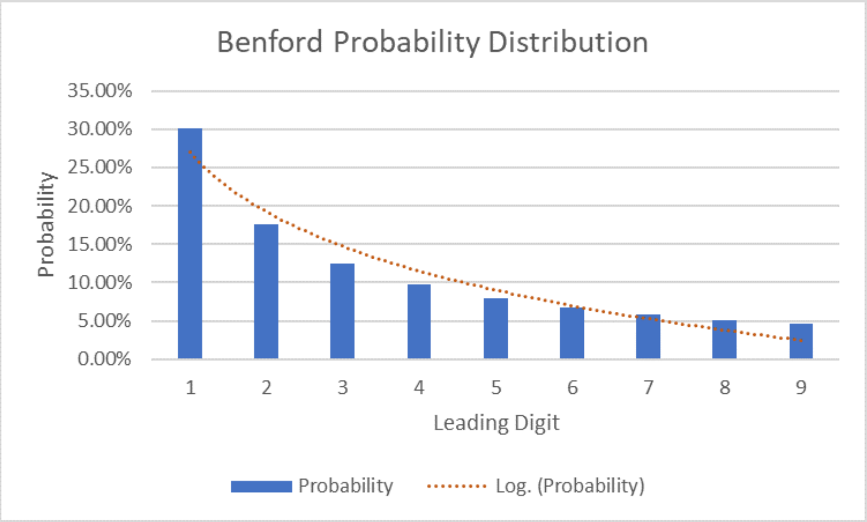
Table 1. The period under investigation by SEBI

a. Statistical Approach – Benford’s Law

Benford's Law is a very simple statistical law that can be used to detect the probability of fraud in any given dataset. Statistically, Benford's law is known as the first-digit law, which states that in many naturally occurring datasets, the first digit of a number is usually small rather than large. In particular, the law predicts that the first digit will appear about 30% of the time, while the digit 9 will appear less than 5% of the time.

If a dataset is significantly different from the expected distribution of the first digits, this may be an indication that it has been altered or fabricated. For example, if a company's financial statements show an unusually high proportion of numbers starting with the digit 9, it could be a red flag for fraudulent activity. In our study, we examined the total number of shares column, which represents the volume of shares traded that day. Our analysis of the dataset using Benford's Law confirms that there has been manipulation.

We know that Benford's Law is a simple statistical law to detect data manipulation, but it simply gives a YES or NO answer, that is, whether manipulation is detected or not. We need precise periods for the manipulation, so we investigated LSTM with autoencoders and TadGANs further in our research. We concluded that these algorithms performed better in manipulation detection and its precise period.



b. LSTM Autoencoder

Nitish et al. have one of the earliest mentions that LSTMs could be enhanced with learning embedding from an encoder-decoder model. They presented the use of multilayer Long Short-Term Memory (LSTM) networks to learn representations of sequences data. For a given dataset of sequences, an encoder-decoder LSTM is configured to read the input sequence, encode it, decode it, and recreate it. The performance of the model is evaluated based on the model’s ability to recreate the input sequence. Once the model achieves a desired level of performance recreating the sequence, the decoder part of the model may be removed, leaving just the encoder model. This model can then be used to encode input sequences to a fixed-length vector. This model design enables the effective processing of sequential data, capturing temporal patterns, and generating the desired output. The model architecture is depicted in Figure 2. The model was trained using the pipeline shown in Figure 1.

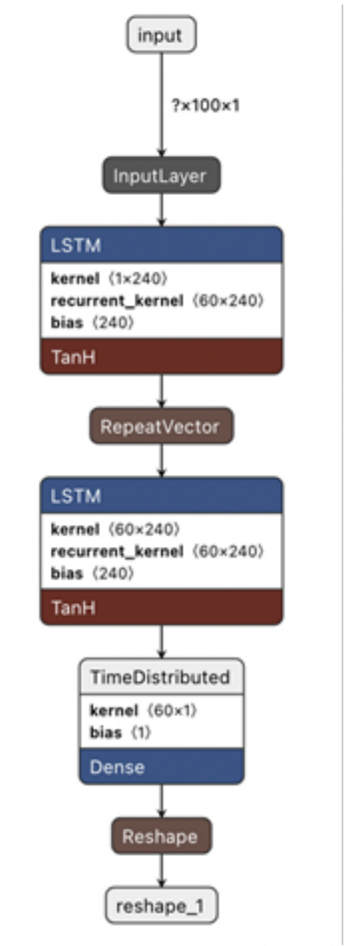


Figure 2. LSTM Autoencoder Architecture

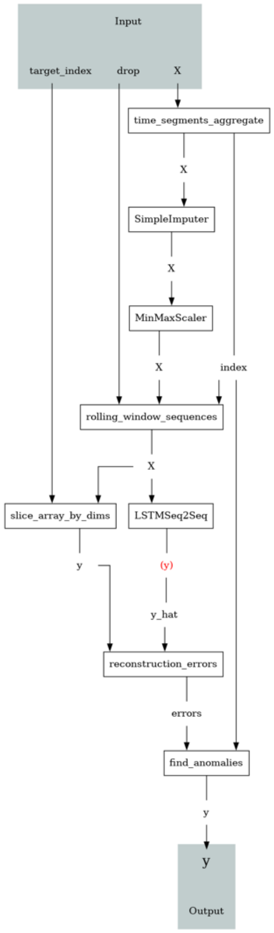


Figure 1. LSTM Autoencoder training pipeline

c. GANs Approach

Mentioned in Liu et al. TadGAN [12] offers a performance-efficient and generalisable approach for anomaly detection. With an adversarial unsupervised learning approach, they can capture temporal correlations of the time series distribution. The original cycle loss method described in the paper allows efficient reconstruction of the time series. To reconstruct signals only Generators and Encoders are used which can be represented as $G(E(s)) \approx \hat{s}$.

The Generators and Encoders instinctively should not be able to reconstruct the anomaly. Henceforth, anomalous stock data should deviate from the reconstructed ŝ. The critic C_x is responsible for identifying what windows are anomalous in $\hat{s}$. The architecture of TadGAN is represented in Figure 3 & Figure 4. The model was trained using the pipeline shown in Figure 5.

The TadGAN model in the research paper was configured with the following statistics: input sequence length of 100, a 20-dimensional latent space, a batch size of 64, a 1-

layer bidirectional LSTM with 100 hidden units for the Encoder, a 2-layer bidirectional LSTM with 64 hidden units for the Generator, a 1-D convolutional layer for Critics, and training for 25 epochs. The default window size of 100 was used for sub-segmenting the stock.

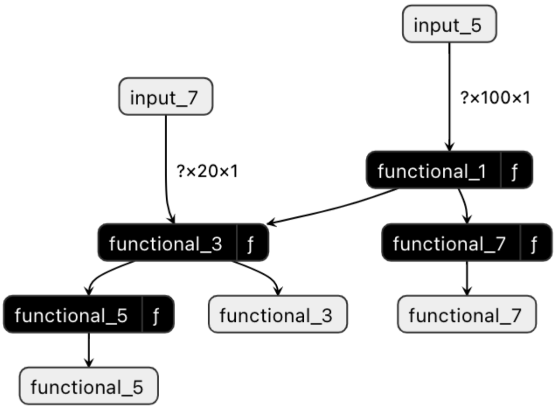


Figure 5. Encoder Generator Model Architecture

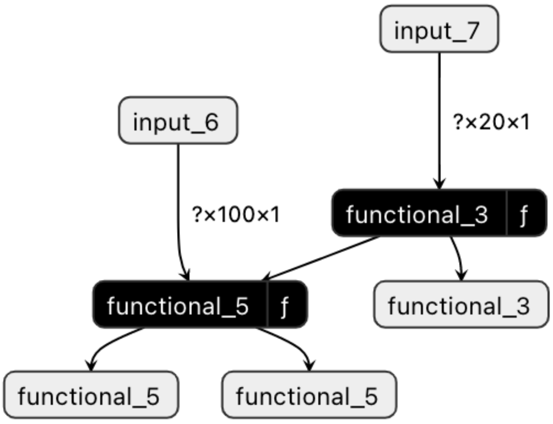


Figure 4. Critic Model Architecture

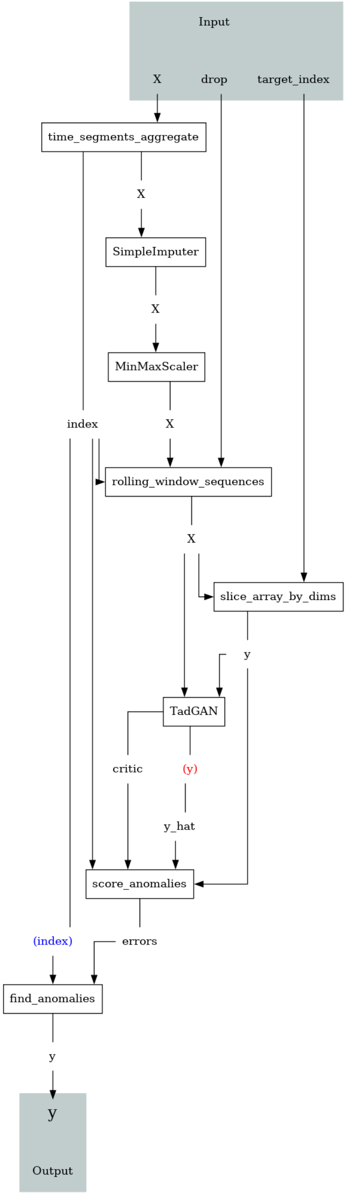
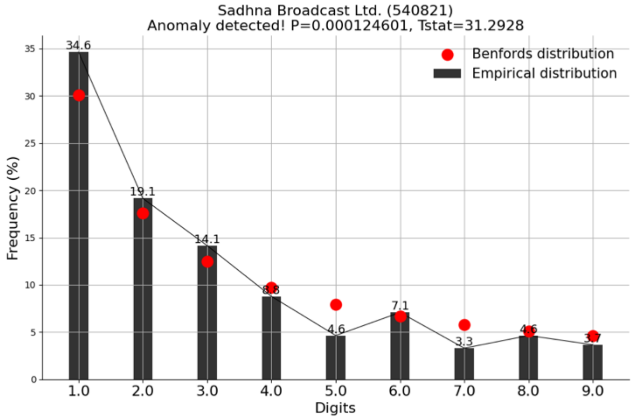
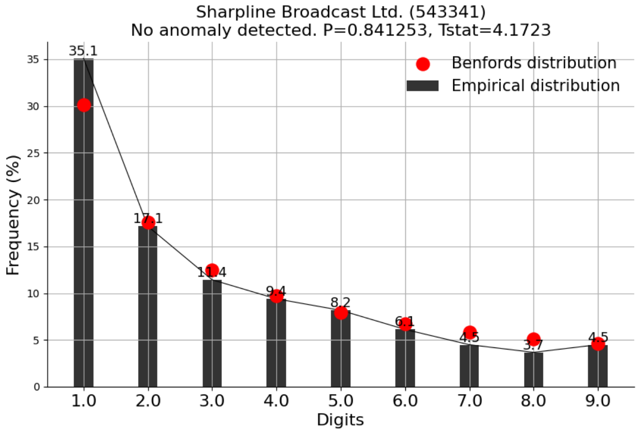


Figure 3. TadGAN training pipeline

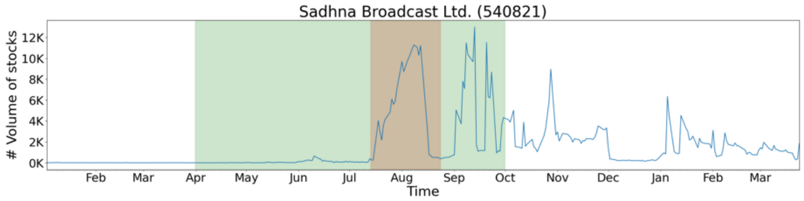
IV. RESULTS & DISCUSSION

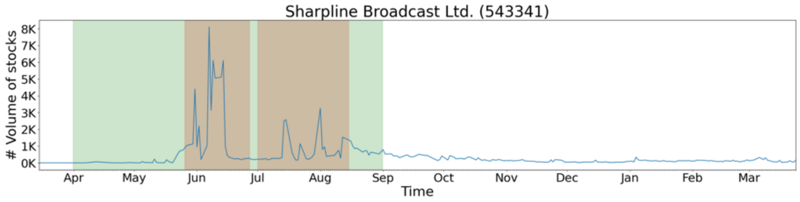
c. Benford's Law





d. LSTM Autoencoder

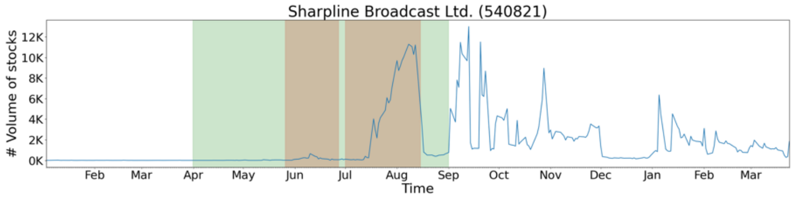


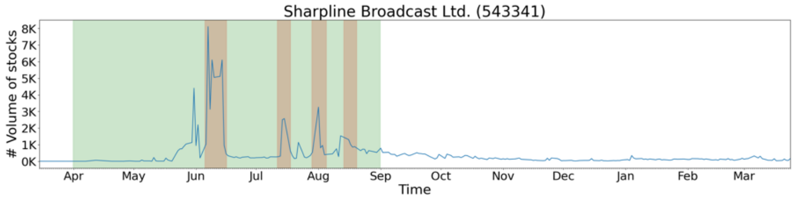


Name	F1 Score	Recall
Sadhna Broadcast Ltd.	0.367893	0.225410
Sharpline Broadcast Ltd.	0.669565	0.503268

Table 2

e. TadGANs





Name	F1 Score	Recall
Sadhna Broadcast Ltd.	0.251244	0.143670
Sharpline Broadcast Ltd.	0.339213	0.204249

Table 3

V. CONCLUSION & FUTURE WORK

The results of the study demonstrate that market manipulation can indeed be detected using these techniques, although the accuracy varies and can be further improved. Among the deep learning techniques, the LSTM with Dynamic Thresholding shows promise in this domain. It effectively identifies contextual and local anomalies in the data and has the advantage of quickly detecting anomalies, with the ability to score two years of trading data for each stock within seconds. This is a notable achievement, considering that deep learning approaches often struggle with processing large volumes of data efficiently.

For future research on the identification of market manipulation, the paper suggests considering hybrid methods that combine both deep learning and statistical techniques. The authors refer to a study by Buda et al. [3], which describes such hybrid approaches. Additionally, the paper acknowledges the strong performance of the ARIMA model in detecting point anomalies and suggests incorporating it into future work.

In the context of the paper, the employed techniques for stock market manipulation detection include Benford's Law, LSTM (Long Short-Term Memory), TAD-GAN (Time Series Anomaly Detection using Generative Adversarial Networks), and time series anomaly detection methods.

VI. ENVIRONMENT

The primary hardware was powered by the Ampere Altra processor, hosted on Oracle Cloud through Ampere A1 compute services. Python v3.8.15 with CPU-optimised Tensorflow v2.3.4 and Orion-ml [13] v0.4.1 was used with processor clock speed @3Ghz and 24Gib RAM.

VII. REFERENCES

[1] Sashank *Sridhar, Siddhartha Mootha & Dr Sudha Subramanian, "Detection of market manipulation using ensemble neural networks.",* International conference on Intelligent Systems and Computer Vision (ISCV), 2020.

[2] T. *Leangarun, P. Tangamchit & S. Thajchayapong, "Stock price manipulation detection using a computational neural network model",* 2016 Eighth International Conference on Advanced Computational Intelligence (ICACI), 2016.

[3] *K. Golmohammadi, O. R. Zaine & D. Diaz, "Detecting stock market manipulation using supervised learning algorithms",* 2014 International Conference on Data Science and Advanced Analytics (DSAA), Shanghai, 2014.

[4] *T. Leangarun, P. Tangamichit & S. Thajchayapong, "Stock Price Manipulation Detection using Generative Adversarial Networks",* 2018 IEEE Symposium Series on Computational Intelligence (SSCI), Bangalore, India, 2018.

[5] *Q.Wang, W.Xu, X.Huang & K.Yang, "Enhancing intraday stock price manipulation detection by leveraging recurrent neural networks with ensemble learning",* Neurocomputing, vol.347, pp. 46-58, 2019.

[6] *Jillian Tallboys, Ye Zhu & Sutharshan Rajasegarar, "Identification of Stock market manipulation with deep learning"*, 17th International Conference on Advanced Data Mining and Applications (ADMA), 2021.

[7] Munir*, M., Chattha, M.A., Dengel, A., Ahmed, S.: A Comparative Analysis of Traditional and Deep Learning-Based Anomaly Detection Methods for Streaming Data. In: 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA).*

[8] *Munir M., Siddiqui, S.A., Dengel, A., Ahmed, S.: DeepAnT: A Deep Learning Approach for Unsupervised Anomaly Detection in Time Series. IEEE Access 7, 1991–2005 (2019).*

[9] *Pang G.,Shen C.,Cao L.,Hengel,A.V.D.:Deep Learning for Anomaly Detection. ACM Computing Surveys.*

[10] *Golmohammadi, K. & Zaiane, O. R. Time series contextual anomaly detection for detecting market manipulation in the stock market. in 2015 IEEE International Conference on Data Science and Advanced Analytics.*

[11] "*Sebi cracks down on stock manipulation via YouTube, bans Arshad Warsi, others from the securities market." The Economic Times, 04 Mar. 2023.*

[12] Geiger*, Alexander and Liu, Dongyu and Alnegheimish, Sarah and Cuesta-Infante, Alfredo and Veeramachaneni, Kalyan "TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks", 2020 IEEE International Conference on Big Data (IEEE BigData)*

[13] *Alnegheimish, Sarah and Liu, Dongyu and Sala, Carles and Berti-Equille, Laure and Veeramachaneni, Kalyan "Sintel: A Machine Learning Framework to Extract Insights from Signals", Proceedings of the 2022 International Conference on Management of Data, 2022.*