Journal of Integrated
SCIENCE & TECHNOLOGY

Outlier detection and imputation of missing data in stock related time series mulitivariate data using LSTM autoencoder

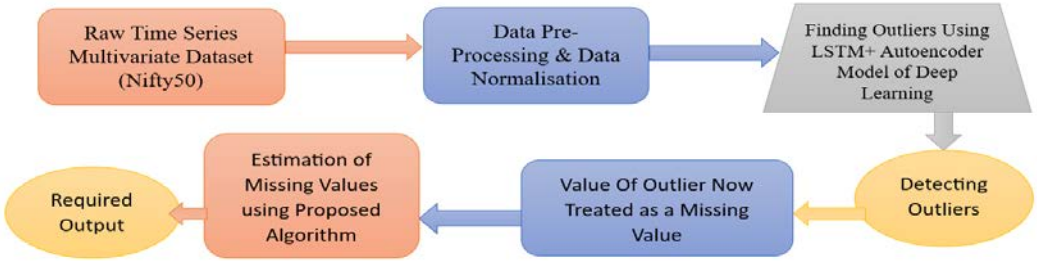
Swati Jain*, Naveen Choudhary, Kalpana Jain

Department of Computer Science and Engineering, College of Technology and Engineering, Udaipur, (Raj), India.

Received on: 17-Sep-2023, Accepted and Published on: 01-Dec-2023

Article

ABSTRACT



Incomplete data is a well-known problem with large databases, which raises challenges for many data mining applications. The main focus will be on developing scalable and adaptable anomaly detection techniques that can spot unusual trade patterns in vast amount of stock related data. In the designed method, an autoencoder based unsupervised outlier detection for multivariate time series data has been trained within a Long Short Term Memory Model, shaped by deep learning networks. Further, we also estimated value of outlier which is treated as missing value in dataset using our designed algorithm. The suggested method demonstrates the value of feature selection and data preparation for creating effective techniques for data modeling. Deep learning models can be tuned very little to produce good results. Our work used the Stock related dataset, that many investors choose due to its high risk, high reward, and flexible trading. In comparison, our designed work also examines the statistics and machine learning models in related application fields.

Keywords: Deep Learning, Feature Extraction, Stock Market, Machine Learning.

INTRODUCTION

Detecting outliers has remained a vibrant field of study for many decades, owing to its wide-ranging applications across critical domains, including risk management, compliance, security, financial surveillance, medical risk assessment, and safety. While extensively explored in fields such as data mining, machine learning and statistics, it still presents unique complexities and challenges necessitating advanced approaches.[1] Currently, it enjoys renewed attention after some years of relative neglect, with recent research successfully addressing long-standing issues in outlier theory.[2] In this research paper, we tackle the challenge of detecting outliers in a multivariate stock market dataset using Autoencoder with Long Short-Term Memory, bridging the gap between technical analysis and deep learning techniques. Neural networks have found applications in various domains, including time series analysis. In our outlier detection dataset, we first train an autoencoder. Subsequently, the encoder extracts feature and feeds them into a model to identify outlier values within the stock index. Our system continues to detect outliers in stock data based on the established model.[3] Additionally, we introduce a novel algorithm to estimate the value of outliers, treating them as missing data points in the dataset. Our evaluation metrics encompass mean absolute error, root mean square error, among others. The outcomes show how much better our suggested way is than other methods.

Detecting outliers in stock prices has always been a focal and challenging research area. Traditional methods such as statistics, may not sufficiently deal with complex and dynamic nature of the stock market.[4] Since the 1970s, researchers have increasingly turned to machine learning to detect stock price anomalies and instabilities.[5] In recent years, earlier models such as Autoregressive Integrated Moving Average have witnessed widespread adoption of machine learning techniques, leading to the development of more suitable models for outlier detection.[6] In particular, Deep Learning has shown to be more efficient than previous models, with neural network models outperforming regression and discriminant methods. Some studies have looked into the relationship between new features and stock prices in regard to feature selection.[7] Nevertheless, earlier outlier detection approaches often relied on

*Corresponding Author: Swati Jain
MPUAT University, CTAE, Udaipur, (Raj), India.
Tel: 7014650144
Email: swati.subhi.9@gmail.com

Cite as: *J. Integr. Sci. Technol., 2024, 12(3), 761.*
URN:NBN:sciencein.jist.2024.v12.761

techniques like Latent Dirichlet Allocation and Principal Component Analysis to reduce feature dimensions, which may not be well-suited for outlier detection.[8]

This study advances the field by offering a novel method for finding outliers in stock price datasets, leveraging deep learning methods, autoencoder and Long short term Memory model to address the aforementioned research gaps.[9] Our study makes primary contributions such as extraction, cleaning of multiple stock-related datasets from the Nifty 50, an open-source data API. A novel approach to outlier detection, specifically tailored to the stock price domain, capitalizing on deep learning techniques.Top of Form Our work employs Long Short-Term Memory[10] autoencoders, focusing on some key aspects such as model architecture, comparision with prior research, missing data estimation and through feature engineering. Model architecture is used with outlier detection and estimating missing values. Comparison with prior research encompasses a comprehensive comparative analysis with existing research, highlighting the uniqueness and effectiveness of our designed methodology. Missing data estimation is a distinctive facet of our research which involves the calculation of estimated values for outliers, which are now treated as missing data points. We have introduced an innovative algorithm designed specifically for locating and estimating missing data. Through feature engineering, we have placed significant emphasis on ensuring a meticulous and comprehensive process to optimize model performance. Our research journey begins with an exhaustive review of prior works, meticulously identifying gaps and challenges in the existing literature.

Drawing from this analysis, we have devised a robust solution architecture, aiming to achieve exceptional detection accuracy across various evaluation metrics. This model is engineered to capture daily features within time series data and subsequently detect abnormalities using the encoded data representation. This novel method offers a substantial breakthrough in the field of outlier detection by improving our capacity to spot outliers and abnormalities within the context of time series data.

DESIGNED MODEL

2.1 Long Short-Term Memory(LSTM) Networks

LSTM networks represent an enhanced iteration of neural networks that excel at capturing prolonged dependencies in data while effectively addressing the vanishing gradient problem inherent in traditional LSTM networks. The crucial element of LSTM are known as the memory cell.[11] Comprising three sigmoid and one hyperbolic tangent layers, this cell forms three gates that intricately manage the flow of information within and beyond its boundaries. The input and output gates govern the inflow and outflow of information within the memory unit, respectively, while the forget gate, employing a sigmoid function, possesses the capability to selectively reset the memory unit.[12]

2.2 Autoencoder

An autoencoder constitutes an artificial neural network characterized by two integral components: an encoder and a decoder. The main objective of autoencoders is to replicate the input data at the networks output.[13] During this process, autoencoders inherently acquire valuable data characteristics and patterns, facilitating a deep understanding of the data. The input sequence needs to be encoded to produce a representation of the original data and then decoded to produce a constructed sequence in order to develop an LSTM auto-encoder. The dimensions of the input data must be reduced by the encoder to create an encoded form. The smallest possible dimensions should be used for this depiction. After receiving this representation, the decoder will try to recreate the original sequence using the same length as the original input. By dimensionality reduction, this will enable the model to discover the fundamental patterns in the data inside the lag. Any events that do not follow the core pattern can be identified because the core pattern was reproduced at the decoder stage's output. As seen in Figure 1,

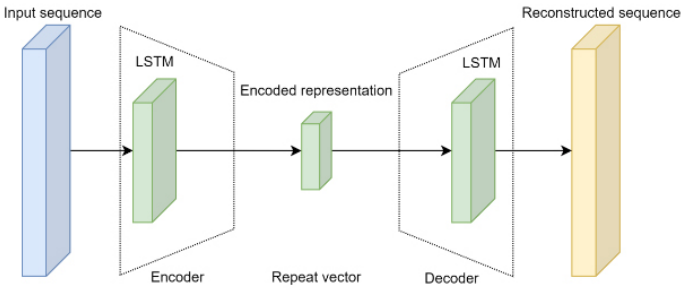


Figure 1: Representation of an LSTM Autoencoder.

2.3 Unsupervised Multivariate Time Series Data

Time series data is a collection of observations systematically recorded at consistent intervals, creating a sequential dataset that effectively captures recurring patterns and phenomena over time.[14] This data category pervades a multitude of domains and applications, including but not limited to monitoring stock prices, heart rate measurements, quantifying rainfall fluctuations, analyzing brain activity patterns, and conducting environmental assessments etc. When we examine the distribution of more than single feature or more single data space to find anomalies is called multivariate outliers. Multivariate data consist of more than one column with timestamp.[15] When a machine learns a function from input features that map to output features using input-output pairs, it is said to be engaging in unsupervised learning. The data is not labeled. More secure learning has Unsupervised learning comparison with supervised learning in the anomaly detection. [16]

2.4 Missing Data

Missing data are undoubtedly those values in the data collection that are either absent for regular or unusual reasons, misplaced, unrealistic, or not supplied at all. Missing values in the data cause confusion for both data analysis and filing of an explanation for new data. The three types of mean imputation techniques introduced for missing data were established by Noor et al.[17] Inference, missing data, and multiple imputations utilized to account for non-response in the study were all topics covered by Rubin.[18] Allison looked into linear model estimates using insufficient data.[19] According to Smyth's measurements, the first step in every data study is data preparation.[20] We develop an

innovative approach in our research to replace the missing variables.

2.5 Designed Model

We designed two models, one for outlier detection and the other for estimation of missing values. In the first model, we took raw time series multivariate stock related dataset taken from the Nifty 50 index of Indian stock market. Further, Data preprocessing and normalization was done and then finding outliers using our deep learning model in Figure1. After finding the outliers at various positions in our dataset, we removed them and treated them as missing values. The Designed algorithm for estimating missing values is given below:

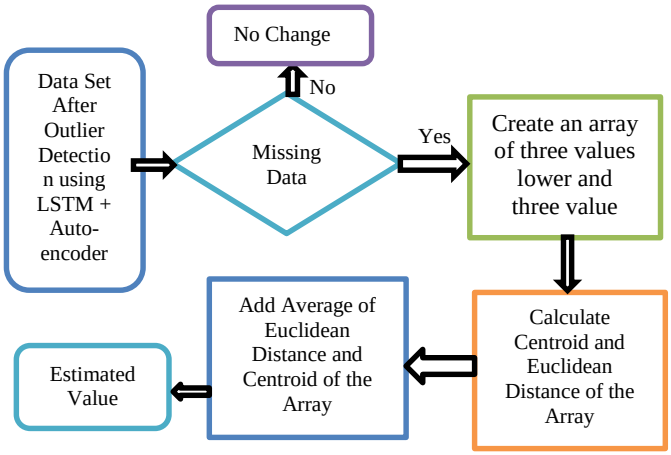


Figure 2: Designed model for estimation of missing value.

The Designed work's workflow is depicted in Figure 2. The new dataset after detecting outlier values which are now treating as missing value contains missing values, a small subset or array is created from the input data sheet. We Create an array of three values lower and three value upper from the missing value.

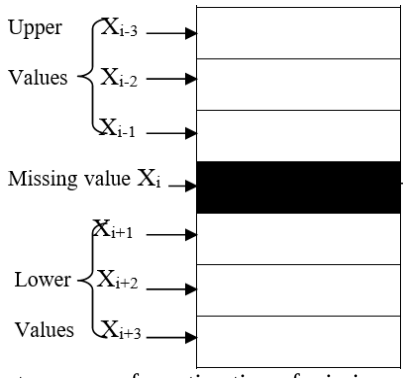


Figure 3: Create an array for estimation of missing value.

Missing value calculation: Estimating missing values during the last stage, after the data is clear of outliers. In order to fill in the missing values in this array, we first compute the centroid of the subset, which is produced by taking the average of the subset. At a later step, the Euclidean distance between each data value and the data centroid is taken into consideration. As shown below, the separation between vectors X and Y is definite:

$$\text{Euclidean distance (d)} = \sqrt{\sum_{i=1}^{n}(\text{Xa} - Ya)}\ 2$$

X_a is centroid of the array, Y_a is particular value of the array

Finally, we add the centroid of the array and average of Euclidean distance. This is the estimated value of the missing item. Every missing value in the entire datasheet has its estimated value, or Xest, assessed one at a time.

X_{est}= Average of Euclidean Distance + Centroid of The Array

2.5 Dataset Description

The Nifty 50 is a diverse 50-stock index that takes into different economic sectors. Nifty 50 is managed by National Stock Exchange Indices Limited (NSE Indices). NSE Indices is a highly specialized company with the index as its primary product. Main columns of stock dataset are as follows: open, high, low, close, volume, date. Open is the market's opening session is known as the open. Low is the lowest price. High is the highest price at which a stock traded. Close refers to the end of a trading session. Volume is the security amount that was traded during a given period of time. We have worked on many stock related Nifty 50 datasets in our research work.

Results And Discussion

Using Keras, the algorithm was built on top of the TensorFlow library. Using a Tesla K80 graphics processing unit (GPU) for hardware acceleration, Anaconda Navigator was used to run the model. 1.416% of the processing time was needed to process the data; 1016.133 s were needed for training the model, and 6.152 s were needed for model execution.

3.1 HDFC Bank dataset

Following is the graph of HDFC NSE stock price data from the year 1996-2020.

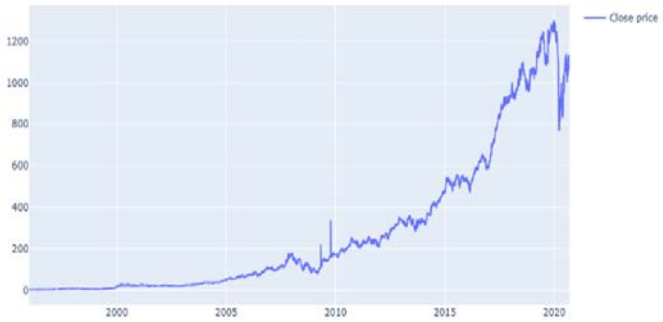


Figure 4: Graphical representation of HDFC bank stock price.

Now we preprocessed the HDFC dataset through the code written in python using pandas, tensorflow, sklearn, numpy, seaborn and matplotlib libraries.

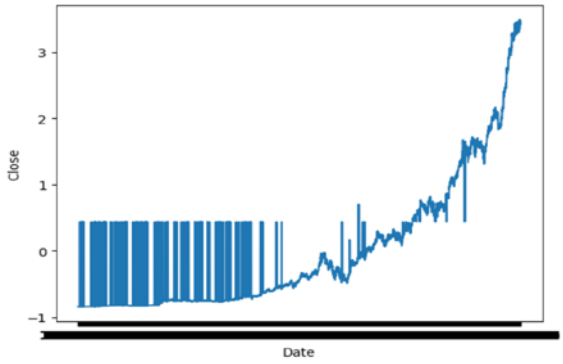


Figure 5: Graphical representation of data preprocessing step in HDFC bank.

In the next step, we divided the above dataset into two parts, namely testing and training dataset. The graphical representation of training and testing loss in our dataset is shown below:

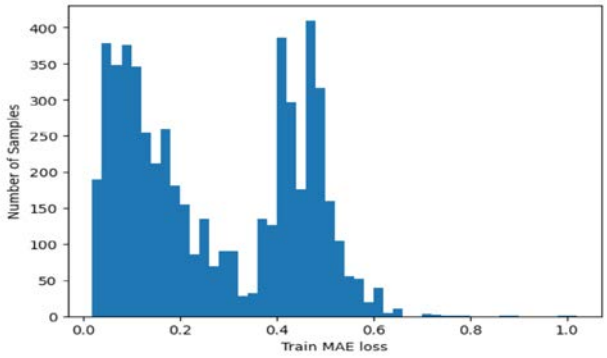


Figure 6: Graphical representation of data training loss.

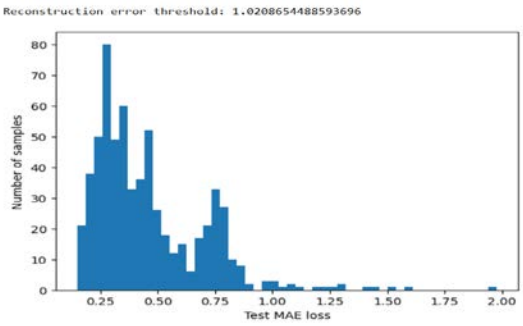


Figure 7: Graphical representation of data testing loss.

Now, plot metrics and evaluate model loss as shown in the graph below:

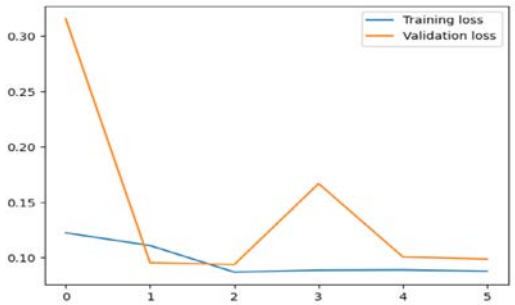


Figure 8: Graphical representation of model loss.

After finding loss, our system generated number of Outliers and their position using Designed model as given below:

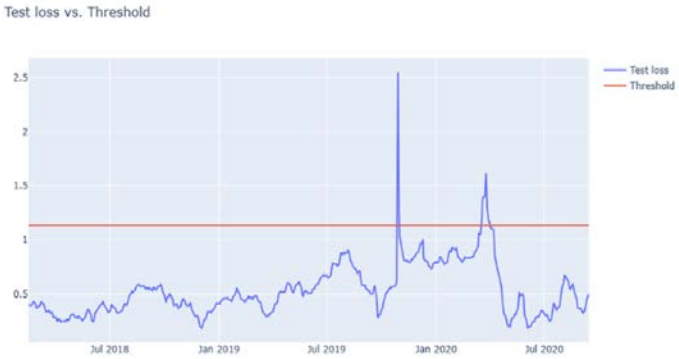


Figure 9: Graph showing outliers in given dataset.

Further, each outlier was treated as missing value which in turn was calculated and estimated using our model. Their locations and estimated missing values are given below:

Table I: Location of outliers and filling estimated value

Locations of Missing Values	After Finding Missing Values at Specified Positions
Row: 6007	1243.33612
Row: 6008	1244.45726
Row: 6009	1245.38112
Row: 6103	1155.15000
Row: 6104	1132.51665
Row: 6105	1102.98888
Row: 6106	1130.21851
Row: 6107	1121.90801
Row: 6108	1118.37180
Row: 6109	1123.49944
Row: 6110	1121.25975
Row: 6111	1121.04366
Row: 6112	1121.93429
Row: 6113	1121.41257
Row: 6114	1121.46351

Comparision of original and Designed datsaset using graph is shown below:

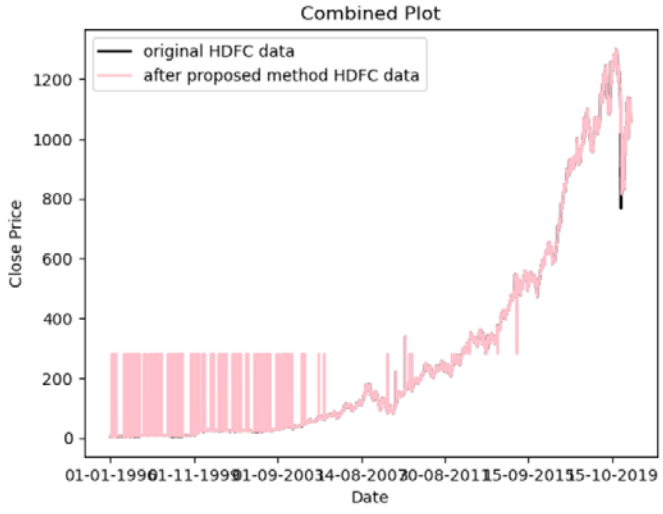


Figure 10: Graphical representation of original and experimental dataset of HDFC.

We followed the same process on Coal India NSE Nifty 50 dataset. Following is the graph of Coal India NSE stock price data from the year 2010-2021.

Figure 11: Graphical representation of Coal India stock price.

Number of Outliers and their position after using Designed model are given below:

Figure 12: Graph showing outliers of given dataset.

Further, each outlier was treated as missing value which in turn was calculated and estimated using our model. Their locations and estimated missing values were calculated. Comparision of original and designed datsaset using graph is shown below:

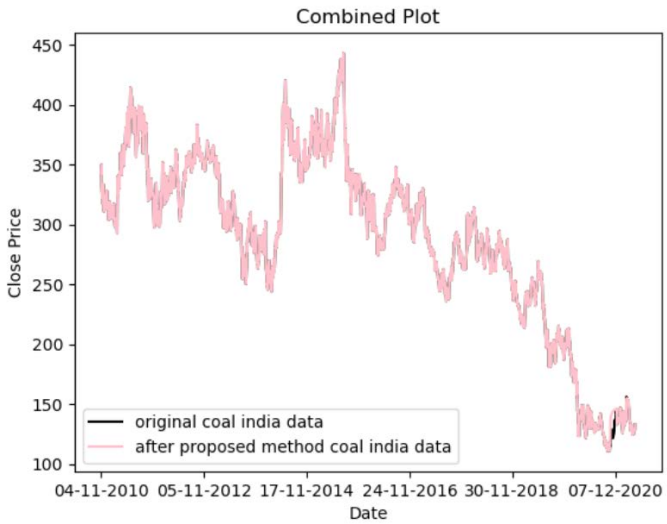


Figure 13: Graphical representation of original and experimental dataset of Coal India.

3.2 Comparitive Study Between Popular Existing Techniques with Our Designed Work:

The Designed method is far better than the other existing techniques for outlier detection in stock related time series multivariate dataset. We examined more than 20 datasets and based on the total number of outliers detected in these datasets, we found our work showed far better results compared to other methods. As we know, a lower total number of detected outliers is generally desirable because it implies that the method is being more conservative i.e. when a method detect fewer outliers, it means it is less likely to flag data points as outliers unless they are genuinely unusual or outliers. Over detecting outliers leads to unnecessary investigation and false results. Thus, a lower total number of detected outliers is generally desirable when it reflects a methods ability to be conservative and accurately identify only the most significant outliers. We compared our designed method with other machine learning and statistical techniques such as SVM, Z- Score and IQR and concluded that our method is more selective and conservative in identifying outliers compared to others.

Table II: Comparitive analysis of designed method with existing methods.

Dataset tested with their rows and columns		Designed Method (LSTM +autoencoder)	ML (one class SVM)	Z-score	IQR
HDFC Bank Dataset (6229 ,7)	Total no of outliers	15	559	528	151
Coal India Dataset (2598,15)	Total no of outliers	37	245	219	294
Maruti Dataset (4093, 15)	Total no of outliers	25	369	235	257
Sun-Pharma Dataset (5059, 15)	Total no of outliers	34	448	490	284

We compute Mean square error, Root Mean Square Error on our datasets for performance evaluation. We use SVM[21], Z-score, and IQR as the standard to evaluate our model because these are established techniques for identifying stock market outliers.[22] The average of the squares of the errors between forecasts and the observed value is measured by the mean squared error (MSE), which gives us an estimate.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$\mathbf{MSE}$ = mean squared error
n = number of data points
Y_i = observed values
$\hat{Y}_i$ = predicted values

The root mean square error, or RMSE, is the standard deviation of the prediction errors. Errors in prediction are the spread of these residuals is determined by RMSE, and residuals are the distance

between the data points and the regression line. Put another way, it shows how much of the data is centered around the line of best fit. The equation is:

$$RMSE = \sqrt{(f-o)^2}$$

In this case, o = observed values (known results), and f = forecasts (anticipated values or unknown results). The mean, or $\bar{x}$, is represented by the bar above the squared differences. With a little variation, the formula can be expressed as follows:

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

Σ = summation ("add up"), $(z_{fi} - Z_{oi})^2$ = differences, squared
N = sample size.

The table displays the prediction outcomes, highlighting the best result. Low Mean Absolute Error and Root Mean Square Error suggest that the closing price predictions closely align with the actual data.[23] Our approach demonstrates competitive performance when compared to other methods.

Table III: Performance Metrics Between Designed and Other Existing Method.

Performance Metrics		Designed Method (LSTM+ autoencoder)	One class SVM (ML)	Z-score	IQR (Inter-Quartile Range)
HDFC Bank Dataset	MSE	0.82323	1.01332	364.459	5.363
	RMSE	0.90732	1.11543	19.090	2.316
Coal India Dataset	MSE	0.91620	1.021555	288.485	5.441
	RMSE	0.95718	1.14681	16.984	2.332
Maruti Dataset	MSE	0.18011	0.61588	3143.751	5.230
	RMSE	0.42439	0.92601	56.069	2.286
Sun Pharma Dataset	MSE	0.10215	1.01304	768.556	5.451
	RMSE	0.31961	1.11421	27.722	2.334

CONCLUSION

In our designed work, we analysed and validate the performance of the designed methodologies in comparison to the popular existing techniques of detecting outliers and filling missing values in multivariate time series stock data with performance matrix such as mean square error, root mean square error that gives more accurate and better results. We have worked on two objectives which are, detecting outliers and generating estimated missing values using our algorithm. We used Nifty 50 stock data and examined more than 20 datasets. After analysis, we got better performance and lower biasness for stock related multivariate time series data. Our method also outperforms known approaches on various data sets, and the results are much lower than those obtained by the validation methods. Our approach has the advantage that it may be applied broadly to any stock because it simply needs to be fine-tuned using identified abnormal periods. It will be necessary to use an entirely unsupervised approach, although more fine-tuning can be done to identify abnormalities that are not immediately apparent. A compensated model is better adapted to handle anomalous periods and modify its prediction to closely match the true value, according to the compensation strategy that was employed. This contrasts with a model who receives no pay. Our hypothesis that it should be possible to identify outlying events and use these detections to improve forecasting is supported by the data. This could involve assessing and refining alternative deep learning techniques. The use of deep learning to identify and correct for abnormalities in financial time series forecasting was suggested in this work. The results corroborate the viability of our suggested approach. Our approach was validated against existing financial time series prediction techniques. Our method produced an error RMSE and MAE. The error indicators employed demonstrate how our methods predictions are regularly in close proximity to the actual number.

FUTURE WORK

Improving LSTM-Autoencoder models involves enhancing their architecture, training methods, and application-specific considerations. Hyperparameter Optimization, Attention Mechanisms, Variational Autoencoder (VAE) Fusion, Regularization Techniques. We can use same method using GAN technique of deep learning networks for outlier detection and missing values of dataset for better performance and lower biasness.

ACKNOWLEDGMENTS

Authors acknowledge the support from MPUAT University, Udaipur for necessary facilities.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. R. Al-amri, R. K. Murugesan, M. Man, A. F. Abdulateef, M. A. Al-Sharafi, A. A. Alkahtani. A review of machine learning and deep learning techniques for anomaly detection in IoT data. *Applied Sciences.* **2021**. 11(12), 5320.
2. G. Pang, C. Shen, L. Cao, A. V. D. Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*. **2021**, 54(2), 1-38.
3. J. Patel, S. Shah, P. Thakkar, K. Kotecha. Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications.* **2015**, 42(1), 259-268.
4. A. N. Refenes, A. Zapranis, G. Francis. Stock performance modeling using neural networks a comparative study with regression models. *Neural Networks*. **1994**, 7(2), 375-388.
5. R. Singh, S. Srivastava. Stock prediction using deep learning. *Multimedia Tools and Applications*. **2017**, 76(18), 18569-18584.
6. Wang X., Wang X., Wilkes, M. Unsupervised fraud detection in environmental time series data. *in New Developments in Unsupervised Outlier Detection*. **2021**, 257-277.
7. X. Wang, M. Wilkes. Unsupervised fraud detection in environmental time series data. In *New Developments in Unsupervised Outlier Detection*. **2021,** 257-277.

8. Z. Ji, J. Gong, J. Feng. A Novel Deep Learning Approach for Anomaly Detection of Time Series Data. *Scientific Programming*, **2021**.
9. M. Said Elsayed, N. A. Le-Khac, S. Dev, A. D. Jurcut. Network anomaly detection using LSTM based autoencoder. In *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*. **2020**, 37-45.
10. S. Ghimire, R. C. Deo, H. Wang, M. S. Al-Musaylh, D. Casillas-Perez, S. Salcedo Sanz. Stacked LSTM Sequence-to-Sequence Autoencoder with Feature Selection for Daily Solar Radiation Prediction: A Review and New Modeling Results. *Energies*. **2022**, 15(3), 1061.
11. P. Malhotra, A. Ramakrishnan, G. Anand, L. Vig, P. Agarwal, G. Shroff. LSTM-based encoder-decoder for multi-sensor anomaly detection. *arXiv preprint arXiv*. **2016**, 1607.00148.
12. S. Hochreiter, J. Schmidhuber. Long short-term memory. *J Neural Comput.* **1997**, 9(8), 1735–80.
13. A. Essien, C. Giannetti. A deep learning framework for univariate time series prediction using convolutional LSTM stacked autoencoders. In *2019 IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA).* **2019**, 1-6.
14. A. Lopez Oriona, J. A. Vilar. Outlier detection for multivariate time series: A functional data approach. *Knowledge Based Systems*, **2021**, *233*, 107527.
15. K. Yadav, M. Yadav, S. Saini. Stock values predictions using deep learning-based hybrid models. *CAAI Transactions on Intelligence Technology*. **2022**, 7(1), 107-116.
16. D. Miljkovic. Review of novelty detection methods. In *The 33rd International Convention MIPRO*. **2010**, 593-598.
17. M.N. Noor, A.S. Yahaya, N.A. Ramli, A.M.M AI Bakri. Mean imputation techniques for filling the missing observations in air pollution dataset. In *Key Engineering Materials*. **2014**, 594, 902-908.
18. D. B. Rubin. Inference and missing data. *Biometrika*. **1976**, 63(3), 581-592.
19. P. D. Allison. Estimation of linear models with incomplete data. *Sociological methodology*. **1987**, 71-103.
20. P. Smyth. Data mining at the interface of computer science and statistics. In *Data mining for scientific and engineering applications.* **2001**, 35-61. Springer US.
21. Y. Kara, M. Acar Boyacioglu, O. K. Baykan . Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the Istanbul Stock Exchange. *Expert Syst Appl*. **2011**, 38(5), 5311–9.
22. R.T.F. Nazario, J.L. e Silva, V.A. Sobreiro, H. Kimura. A literature review of technical analysis on stock markets. *The Quarterly Review of Economics and Finance*. **2017**, 66, 115–126.
23. E. Panjei, L. Gruenwald, E. Leal, C. Nguyen, S. Silvia. A survey on outlier explanations. *The VLDB Journal*, **2022**, 1-32.