Pattern Discovery and Prediction in Stock Indices: A TPA-LSTM Multivariate Time Series Analysis

K. Hari Priya[1], P. Vinitha[2], Madhuri Avanthi[2], Mamillapally Sai Anvesh[2], P. Bhargav Naidu[2], P. Raghupatruni Sai Ganesh[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of Information Technology
[1,2]Malla Reddy Engineering College and Management Sciences, Medchal, Hyderabad

ABSTRACT

Forecasting stock index is a crucial aspect in financial time series prediction. Investors have the option to engage in passive investing by investing in a stock index, or they can evaluate the success of active investments by comparing them to a stock index. Hence, it is crucial for investors and professional analysts to establish a more accurate model for forecasting stock indices. Nevertheless, the properties of stock indices, such as being "noisy" and "non-stationary", pose obstacles for prediction. The term "noisy" suggests that there is a lack of sufficient data for investors to accurately assess the historical performance of the stock index. The term 'nonstationary' refers to the characteristic of a stock index to undergo significant changes across different time periods. The aforementioned attributes result in subpar outcomes when forecasting stock index movements using conventional econometric models like the linear model, Auto-Regressive Integrated Moving Average (ARIMA), and Vector Auto Regression (VAR). The aforementioned methods pertain to short-term forecasts in time series, which are significantly influenced by "noisy" and "non-stationary" factors. However, by solely concentrating on predicting the trajectory of stock indices within a specific timeframe, the influence of unpredictable and fluctuating factors on the accuracy of the predictions would be eliminated. One method for predicting the trend of a stock index over a specific period is to break down the long-term stock index into multiple short-term stock index fragments. These short-term fragments are then classified into various patterns, such as "W" shape or "head and shoulders", within pattern sets. The essence of the aforementioned approach is to identify intrinsic patterns within the current stock index sequences and accurately forecast them, commonly known as "pattern discovery" and "pattern prediction". Based on the identified recurring patterns from earlier research, we can anticipate the repetition of stock index patterns within a specific future timeframe. This enables us to strategically take steps to maximize profits and minimize losses. Hence, this article is dedicated to uncovering and predicting patterns in stock indices. This research aims to uncover and forecast trends in stock indices by analyzing multivariate time series data. Our motivation stems from the belief that utilizing pattern recognition and predictive analysis of stock index prices may offer a more practical and efficient approach to financial planning compared to conventional methods like the Autoregressive Integrated Moving Average (ARIMA) model. The application of a three-stage architecture, which combines Temporal Pattern Attention and Long-Short-Term Memory (TPA-LSTM), is utilized for the purpose of discovering and predicting patterns in stock indices.

Keywords: Stock Index Pattern, TPA-LSTM, Vextor Auto Regression.

1. INTRODUCTION

Stock index prediction is one of the most important subjects in financial time series forecasting. Investors can invest passively through a stock index or compare active investment performance with stock index. Therefore, developing a more realistic model to predict stock index is of great importance for investors and professional analysts. However, stock index characteristics, including ‘‘noisy’’ and ‘‘non-stationary’’, make prediction face challenges. ‘‘Noisy’’ implies that there is insufficient information for investors to observe past behaviors of stock index. ‘Nonstationary’ means that stock

index may change dramatically in different periods. These characteristics lead to poor stock index prediction results as predicted by traditional econometric models such as linear model, Auto-Regressive Integrated Moving Average (ARIMA), and Vector Auto Regression (VAR) [1]. The aforementioned methods belong to short-term predictions in time series, which are seriously affected by ‘‘noisy’’ and ‘‘non-stationary’’. However, if stock index prediction only focuses on forecasting the trend over a certain period, the effects of ‘‘noisy’’ and ‘‘non-stationary’’ on the prediction results will be eliminated. One of methods of stock index trend prediction over a certain period is to decompose stock index over long-period into many stock index fragments over short-period, and then stock index fragments over short period are classified into some pattern in pattern sets including ‘‘W’’ shape, ‘‘head and shoulders’’, etc. The core of the process above is to find some inherent patterns in the existing stock index sequences and make appropriate prediction, which is referred to as ‘‘pattern discovery’’ and ‘‘pattern prediction’’. According to the repeated patterns which were discovered in previous work, we can predict repeated patterns of stock index over a certain period in the future, and then make appropriate actions to take gains and avoid losses more effectively. Therefore, this paper focuses on discovering and forecasting stock index patterns.

In this study, we attempt to discover and predict stock index pattern through a three-stage architecture that consists of Toeplitz Inverse Covariance-Based Clustering (TICC), Temporal Pattern Attention and Long- Short Term Memory (TPA-LSTM) and Multivariate LSTM-FCNs (MLSTM-FCN, MALSTM-FCN), which are developed by Hallac et al. [2], Shih et al. [3] and Karim et al. [4], respectively. Taking Hangseng Composite Stock Index (HSCI) and 11 industry stock indices in HSCI as an example, this paper investigates the feasibility of proposed three-stage architecture in financial time series. In the first stage, this paper applies TICC algorithm to cluster prices of industrial indices in HSCI, including consumer good manufacturing, consumer service, energy, finance, industry, information technology, integrated industry, raw material, real estate, utilities. Then, this paper maps the clustering results of industry indices to HSCI and discover repeated patterns of HSCI. In the second stage, TPA-LSTM is used to predict industry indices. In the third stage, this paper applies Multivariate LSTM-FCNs to classify industry indices and predict the pattern of HSCI in the future. Based on the idea that industry indicators are predominant factors in explaining stock market co-movements [5], the proposed three-stage architecture with TICC, TPA-LSTM and Multivariate LSTM-FCNs might be more effective in discovery and prediction of stock index patterns. Moreover, we could conduct early warnings of stock index and make corresponding measures more efficiently.

2. LITERATURE SURVEY

Ouyang et al. [6] attempted to discover and predict stock index patterns through analysis of multivariate time series. This motivation is based on the notion that financial planning guided by pattern discovery and prediction of stock index prices maybe more realistic and effective than traditional approaches, such as Autoregressive Integrated Moving Average (ARIMA) model. A three-stage architecture constructed by combining Toeplitz Inverse Covariance-Based Clustering (TICC), Temporal Pattern Attention and Long-Short-Term Memory (TPA-LSTM) and Multivariate LSTM-FCNs (MLSTM-FCN and MALSTM-FCN) is applied for pattern discovery and prediction of stock index. In the first stage, this paper used TICC to discover repeated patterns of stock index. Then, in the second stage, TPA-LSTM that considered weak periodic patterns and long short-term information is used to predict multivariate stock indices. Finally, in the third stage, MALSTM-FCN is applied to predict stock index price pattern. The Hangseng Stock Index and eleven industrial sub-indices are used in the experiment.

Wei et al. [7] applied a different approach, Temporal Pattern Attention and Long Short-Term Memory (TPA-LSTM), to predict stock indexes’ prices in different industries included in the Hangseng

Composite Index. TPA-LSTM method is a new prediction model that enables the prediction of multivariate time series simultaneously with a top concern of weak periodic patterns and a mixture of linear and nonlinear structures. Further, the TPA-LSTM method comprises four components, the Temporal Pattern Attention component, the RNN component, and the Autoregressive component. The experiment results indicate that by combining the strengths of convolutional network, recurrent network, temporal attention component, and autoregressive component, the TPA-LSTM method significantly improves state-of-the-art results in multivariate time series forecasting on the dataset of industry stock index prices. With the empirical results, this paper shows that the applied TPA-LSTM method is a satisfactory alternative for multivariate time series forecasting in stock indices.

Sayavong et al. [8] proposed a stock price prediction model based on convolution neural network, which has obvious self-adaptability and self-learning ability. Combining the characteristics of CNN (Convolution Neural Network) and Thai stock market, the data set is trained and tested after pretreatment. On this basis, three stocks (BBL, CAPLL&PTT) listed on the Thai Stock Exchange are tested and compared with the actual stock price. The results showed that the model based on CNN can effectively identify the changing trend of stock price and predict it which can provide valuable reference for stock price forecast. The prediction accuracy is high, and it is worth further promotion in the financial field.

Wen et al. [9] introduced a new method to simplify noisy-filled financial temporal series via sequence reconstruction by leveraging motifs (frequent patterns), and then utilized a convolutional neural network to capture spatial structure of time series. The experimental results showed the efficiency of this proposed method in feature learning and outperformance with 4%–7% accuracy improvement compared with the traditional signal process methods and frequency trading patterns modeling approach with deep learning in stock trend prediction.

Thakkar et al. [10] conducted a systematic approach to present a survey for the years 2011–2020 by considering articles that have used fusion techniques for various stock market applications and broadly categorize them into information fusion, feature fusion, and model fusion. The major applications of stock market include stock price and trend prediction, risk analysis and return forecasting, index prediction, as well as portfolio management. This work also provided an infographic overview of fusion in stock market prediction and extend this survey for other finely addressed financial prediction problems. Based on the surveyed articles, this paper provided potential future directions and concluding remarks on the significance of applying fusion in stock market.

Dattatray et al. [11] presented the detailed review of 50 research papers suggesting the methodologies, like Bayesian model, Fuzzy classifier, Artificial Neural Networks (ANN), Support Vector Machine (SVM) classifier, Neural Network (NN), Machine Learning Methods and so on, based on stock market prediction. The obtained papers are classified based on different prediction and clustering techniques. The research gaps and the challenges faced by the existing techniques are listed and elaborated, which help the researchers to upgrade the future works. The works are analyzed using certain datasets, software tools, performance evaluation measures, prediction techniques utilized, and performance attained by different techniques. The commonly used technique for attaining effective stock market prediction is ANN and the fuzzy-based technique. Even though a lot of research efforts, the current stock market prediction technique still has many limits. From this survey, it can be concluded that the stock market prediction is a very complex task, and different factors should be considered for predicting the future of the market more accurately and efficiently.

Chung et al. [12] focused on the optimization of feature extraction part of CNN, because this is the most important part of the computational procedure of CNN. This study proposed a method to systematically optimize the parameters for the CNN model by using genetic algorithm (GA). To

verify the effectiveness of this model, this work compared the prediction result with standard artificial neural networks (ANNs) and CNN models. The experimental results showed that the GA-CNN outperforms the comparative models and demonstrated the effectiveness of the hybrid approach of GA and CNN.

3. EXISTING SYSTEM

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

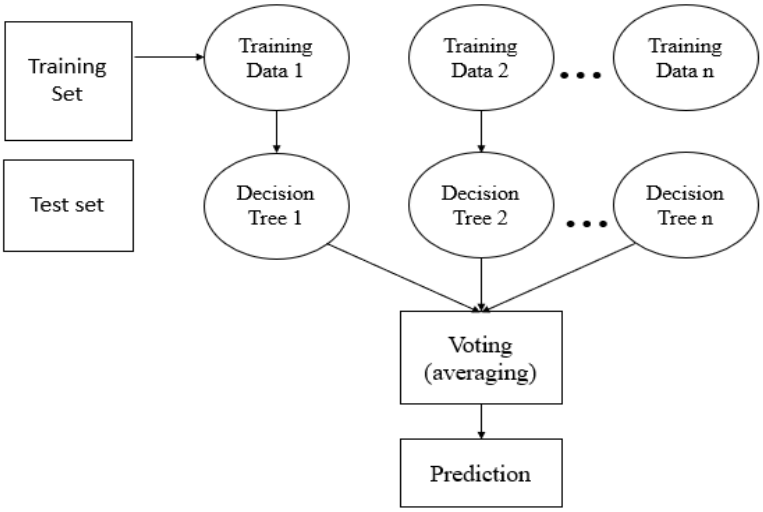


Fig. 1: Random Forest algorithm.

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

Below are some points that explain why we should use the Random Forest algorithm

- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

:**Advantages of Random Forest**

- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.
- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

Applications of Random Forest

There are mainly four sectors where Random Forest mostly used:

- Banking: Banking sector mostly uses this algorithm for the identification of loan risk.
- Medicine: With the help of this algorithm, disease trends and risks of the disease scan be identified.
- Land Use: We can identify the areas of similar land use by this algorithm.
- Marketing: Marketing trends can be identified using this algorithm.

Disadvantages of support vector machine:

- Support vector machine algorithm is not acceptable for large data sets.
- It does not execute very well when the data set has more sound i.e. target classes are overlapping.
- In cases where the number of properties for each data point outstrips the number of training data specimens, the support vector machine will underperform.
- As the support vector classifier works by placing data points, above and below the classifying hyperplane there is no probabilistic clarification for the classification.

2. PROPOSED SYSTEM

The block diagram of the proposed system is shown in Fig. 4.1

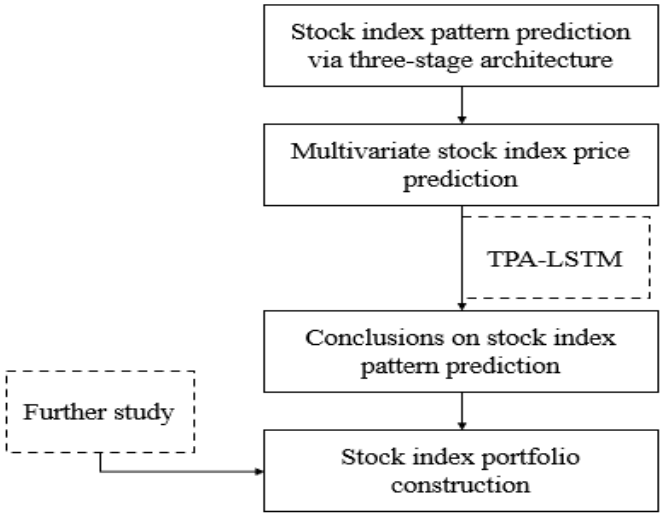


Fig. 2: Block diagram of proposed system.

Dataset description

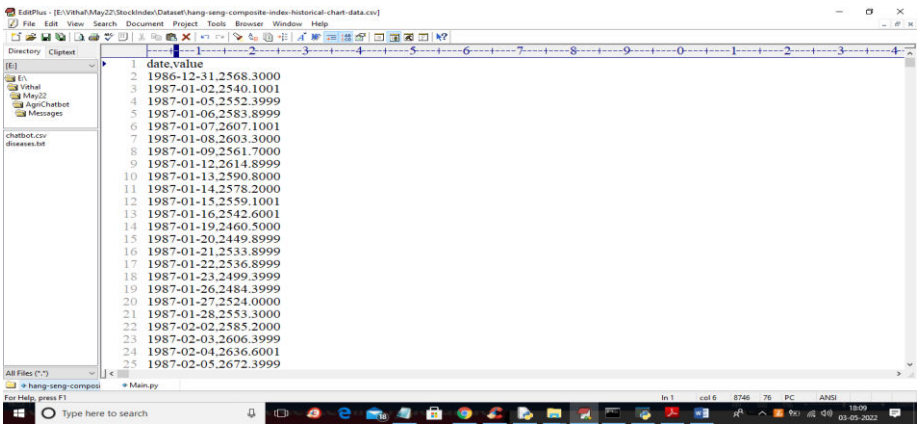
Discovery and Prediction of Stock Index Pattern via Three-Stage Architecture of TICC, TPA-LSTM and Multivariate LSTM-FCNs

In this paper to predict stock index author is using multivariate time series (data which contains time information) data and then extracting repeated values from that data by using Toeplitz Inverse Covariance-Based Clustering (TICC) as this algorithm put similar data in same cluster. Patterns will be extracted by using Temporal Pattern Attention and Long-Short-Term Memory (TPA-LSTM) and then extracted pattern will get trained with Multivariate LSTM-FCNs (fully connected network) to predict stock index.

All existing algorithms such as ARIMA work on time series data but not extract any patterns so its prediction accuracy will be low and Relative Absolute Error will be high.

Author compares propose algorithm TPA-LSTM with various existing algorithms such as SVM, Random forest and Naïve Bayes and evaluate their performance in terms of accuracy and RAE.

To implement this project author has used Hang-Sang dataset and I am also using same dataset and below screen showing dataset details.



In above dataset we have date and stock value and by using this time series and stock data we will train algorithms and then predict stock index and then find difference between original stock index and predicted index as RAE error.

Pre-processing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task.

Why do we need Data Pre-processing?

A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Splitting the Dataset into the Training set and Test set

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model.

Supposeif we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

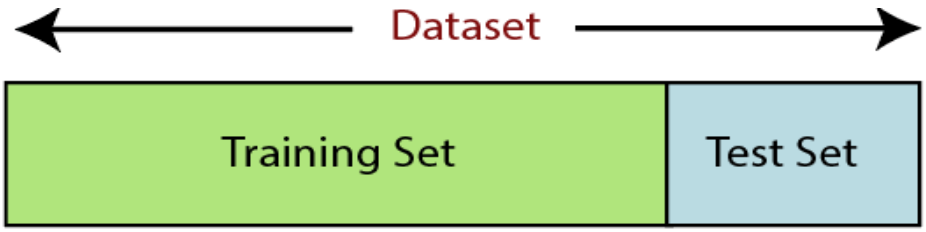


Fig. 3: Splitting the dataset.

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

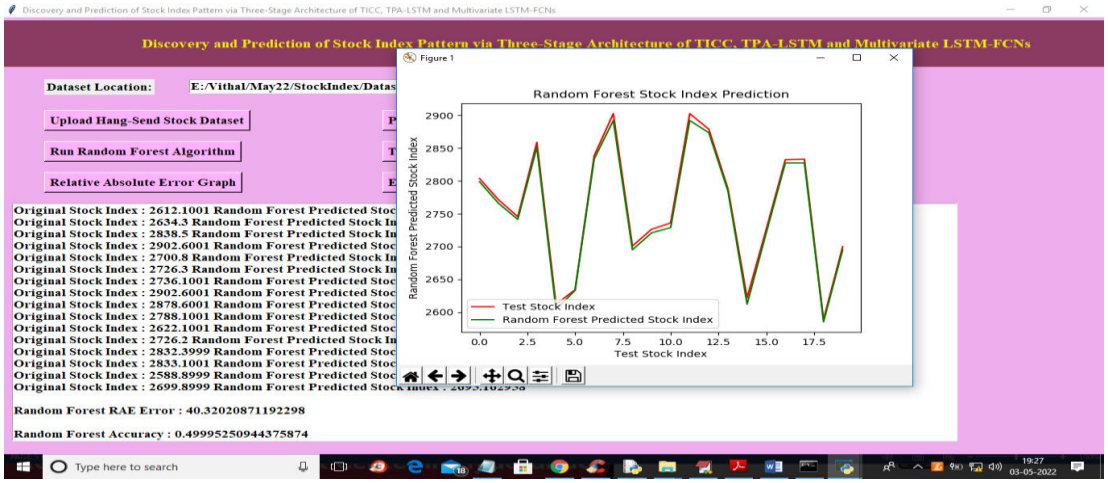
5. RESULTS AND DISCUSSIONS

To implement this project, we have designed following modules.

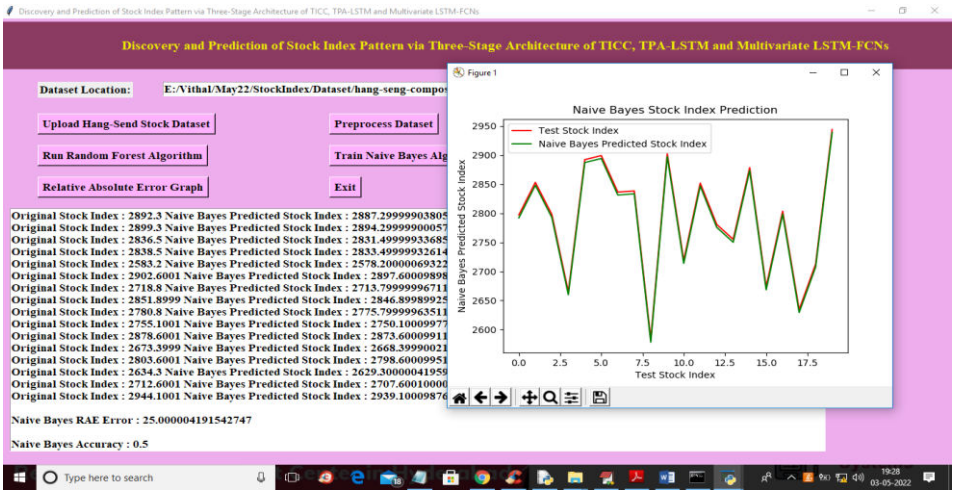
1) Upload Hang-Send Stock Dataset: using this module we will upload dataset to application.
2) Preprocess Dataset: using this module we will read all dataset values and then normalize values using MIN-MAX scaler.
3) Run SVM Algorithm: using this module we will split dataset into train and test and then train SVM on training dataset and then calculate accuracy and RAE on test data prediction.
4) Run Random Forest Algorithm: using this module we will split dataset into train and test and then train Random Forest on training dataset and then calculate accuracy and RAE on test data prediction.
5) Train Naive Bayes Algorithm: using this module we will split dataset into train and test and then train Naïve Bayes on training dataset and then calculate accuracy and RAE on test data prediction.
6) Run Propose TPA-LSTM: using this module we will split dataset into train and test and then train TPA-LSTM on training dataset and then calculate accuracy and RAE on test data prediction.
7) Relative Absolute Error Graph: using this module we will plot RAE (relative absolute error) graph between all algorithms.

SCREENSHOTS

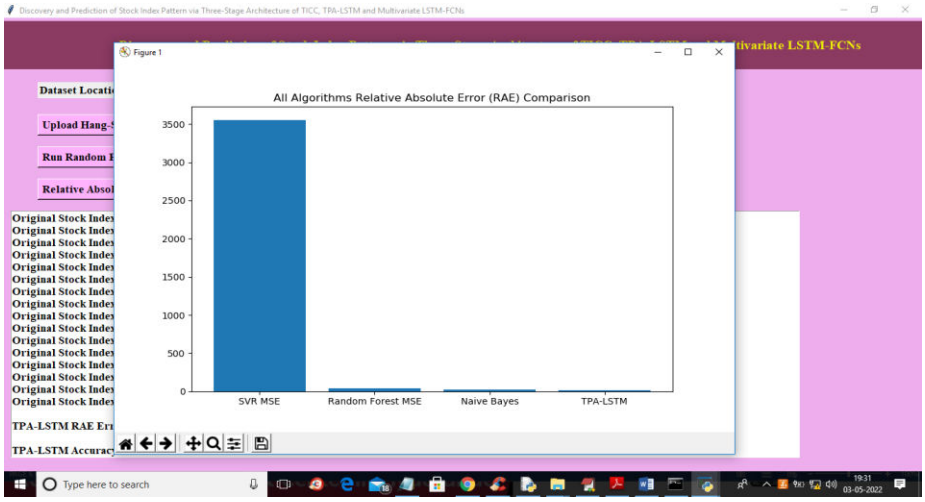
To run project double click on 'run.bat' file to get below screen.



In above screen with random forest both lines are overlapping so its RAE error reduce to 40 and its prediction is little accurate and now close above graph and then click on 'Run Naïve Bayes Algorithm' button to get below output



In above screen with Naïve Bayes we got 25% error rate and both lines are overlapping so its prediction is also little accurate and now close above graph and then click on 'Run Propose TPA-LSTM' button to train propose algorithm and get below output



In above screen x-axis represents algorithm names and y-axis represents RAE error and, in all algorithms, propose LSTM-TPA got less error rate, so its performance is good.

6. CONCLUSION AND FUTURE SCOPE

Discovery and prediction of stock index pattern are of great importance to reduce uncertainty and risks in financial markets and, more specifically, is crucial in constructing a financial portfolio. In the literature of stock index pattern discovery and prediction through neural networks, previous studies mainly focus on pattern discovery and up-down prediction of stock index with strong repeated patterns and fixed time periods. This paper makes up for the shortcomings of previous research, which forms a complete structure of stock index pattern discovery and prediction through a proposed three stage architecture of TICC, TPA-LSTM, and Multivariate LSTM-FCNs. Through proposed three-stage architecture, this paper could analyze and predict stock index prices with weak periodic and flexible patterns.

The proposed three-stage architecture contains three stages. In the first stage, we apply TICC to cluster industry stock indices in the comprehensive stock index and map cluster results to that stock index. Based on the mapping results, we could discover repeated patterns of the comprehensive stock index on the training dataset. In the second stage, we predict multivariate time series of industry stock indices simultaneously through TPA-LSTM. In the third section, we predict repeated patterns of the

comprehensive stock index on the test dataset through Multivariate LSTM-FCNs. HSCI and eleven industry indices that are included in the HSCI are used in the experiment. The empirical results show that the proposed three-stage architecture, including TICC, TPA-LSTM, and Multivariate LSTM-FCNs significantly improves the state-ofthe-art results in pattern discovery and prediction of HSCI. Moreover, we propose a bullish trading rule and construct an equal proportion portfolio based on this trading rule and the prediction results of the proposed three-stage architecture. Seven comprehensive stock indices are used in the experiment. The empirical results show that, the constructed portfolio based on the bullish trading rule and the proposed three-stage architecture performs significantly better than the market-based portfolio. Therefore, the proposed three-stage architecture is a feasible and promising method to discover and predict repeated patterns of stock index in financial markets. There are two promising extensions of pattern discovery and prediction in stock index prices. One possible extension of stock index pattern prediction is to conduct proactive index tracking or construct other trading strategies with predicted patterns of stock index. The other extension is to search for more suitable and effective price prediction and pattern matching methods to improve the performance of the proposed structure.

REFERENCES

[1] Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, ‘‘Support vector regression with chaos-based firefly algorithm for stock market price forecasting,’’ Appl. Soft Comput., vol. 13, no. 2, pp. 947–958, Feb. 2013, doi: 10.1016/j.asoc.2012.09.024.

[2] D. Hallac, S. Vare, S. Boyd, and J. Leskovec, ‘‘Toeplitz inverse covariance based clustering of multivariate time series data,’’ in Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Halifax, NS, Canada, Aug. 2017, pp. 215–223.

[3] S.-Y. Shih, F.-K. Sun, and H.-Y. Lee, ‘‘Temporal pattern attention for multivariate time series forecasting,’’ Mach. Learn., vol. 108, nos. 8–9, pp. 1421–1441, Sep. 2019, doi: 10.1007/s10994-019-05815-0.

[4] F. Karim, S. Majumdar, H. Darabi, and S. Harford, ‘‘Multivariate LSTMFCNs for time series classification,’’ Neural Netw., vol. 116, pp. 237–245, Aug. 2019, doi: 10.1016/j.neunet.2019.04.014.

[5] D. Lamponi, ‘‘Is industry classification useful to predict US Stock Price co-movements?’’ J. Wealth Manage., vol. 17, no. 1, pp. 71–77, Apr. 2014.

[6] Ouyang, Hongbing & Wei, Xiaolu & Wu, Qiufeng. (2020). Discovery and Prediction of Stock Index Pattern via three-stage architecture of TICC, TPA-LSTM and Multivariate LSTM-FCNs (February 2019). IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3005994.

[7] X. Wei, B. Lei, H. Ouyang, Q. Wu, "Stock Index Prices Prediction via Temporal Pattern Attention and Long-Short-Term Memory", Advances in Multimedia, vol. 2020, Article ID 8831893, 7 pages, 2020. https://doi.org/10.1155/2020/8831893

[8] L. Sayavong, Z. Wu and S. Chalita, "Research on Stock Price Prediction Method Based on Convolutional Neural Network," 2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS), Jishou, China, 2019, pp. 173-176, doi: 10.1109/ICVRIS.2019.00050.

[9] M. Wen, P. Li, L. Zhang, and Y. Chen, "Stock Market Trend Prediction Using High-Order Information of Time Series," in IEEE Access, vol. 7, pp. 28299-28308, 2019, doi: 10.1109/ACCESS.2019.2901842.

[10] A. Thakkar, K. Chaudhari, Fusion in stock market prediction: A decade survey on the necessity, recent developments, and potential future directions, Information Fusion, Vol. 65, 2021, Pages 95-107, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2020.08.019.

[11] Dattatray P. Gandhmal, K. Kumar, Systematic analysis and review of stock market prediction techniques, Computer Science Review, Volume 34, 2019, 100190, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2019.08.001.

[12] Chung, H., Shin, Ks. Genetic algorithm-optimized multi-channel convolutional neural network for stock market prediction. Neural Comput & Applic 32, 7897–7914 (2020). https://doi.org/10.1007/s00521-019-04236-3