International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

# IntelliFin: Advanced Stock Prediction using Hybrid ML and LSTM Model with Financial Indicators powered by Sentiment Determination using NLP

**Shashank Singh, Maaz Ahmad, Aditya Bhattacharya, D. Prabhu**

***Abstract*— Stock Trading has been one of the most important parts of the financial world for decades. People investing in the share market analyze the financial history of a corporation, the news related to it and study huge amounts of data so as to predict its stock price trend. The right investment i.e. buying and selling a company stock at the right time leads to monetary benefits and can make one a millionaire overnight. The stock market is an extremely fluctuating platform wherein data is produced in humongous quantities and is influenced by numerous disparate factors such as socio-political issues, financial activities like splits and dividends, news as well as rumors. This work proposes a novel system "IntelliFin" to predict the share market trend. The system uses the various stock market technical indicators along with the company's historical market data trends to predict the share prices. The system employs the sentiment determination of a company's financial and socio-political news for a more accurate prediction. This system is implemented using two models. The first is a hybrid LSTM model optimized by an ADAM optimizer. The other is a hybrid ML model which integrates a Support Vector Regressor, K-Nearest Neighbor classifier, an RF classifier and a Linear Regressor using a Majority Voting algorithm. Both models employ a sentiment analyzer to account for the news impacting the stock prices which is powered by NLP. The models are trained continuously using Reinforcement Learning implemented by the Q-Learning Algorithm to increase the consistency and accuracy. The project aims to support the inexperienced investors, who don't have enough experience in investing in the stock market and help them maximize their profit and minimize or eliminate the losses. The developed system will also serve as a tool for professional investors to help and aid their decision making.***

***Keywords*— LSTM, Support Vector Regressor, K-Nearest Neighbor Classifier, RF Classifier, Sentiment Determination, NLP, Linear Regression, Reinforcement Learning**

Revised Manuscript Received on May 25, 2020.

* Correspondence Author

**Shashank Singh***, Department of Electrical Computer Science and Engineering from SRM Institute of Science and Technology, Chennai.

**Maaz Ahmad,** Department of Electrical Computer Science and Engineering from SRM Institute of Science and Technology, Chennai.

**Aditya Bhattacharya,** Department of Electrical Computer Science and Engineering from SRM Institute of Science and Technology, Chennai.

**D. Prabhu,** Assistant Professor in Computer Science and Engineering at SRM Institute of Science and Technology, Chennai.

## I. INTRODUCTION

The Share market happens to be an erratic system in which copious data is produced at frequent intervals and fluctuates rapidly due to several disparate factors. Investing in the stock market is an efficient way to boost the net worth of an individual. The stock market fluctuates erratically as it is influenced by disparate factors. Investment at the right stock at the correct time leads to monetary benefits. By accurately predicting the share market trend, profits are made relative to the investment. Artificial Intelligence (AI) is widely applied in the field of finance. Techniques like Machine Learning (ML) and Deep Learning (DL) find wide application in the financial domain and in the stock market. Today, the large quantities of stock market data are collated and organized systematically using techniques such as Big Data Analysis and Data Mining. After successfully collating, organizing and visualizing the stock data, the various ML and DL algorithms are used to derive a meaningful inference from that data and also to predict the future stock trends. ML algorithms such as the RF classification, Linear regression, Polynomial Regression, KNN, and SVM have been extensively utilized to predict the stock trends. The DL algorithms like RNN are also applied to the stock market data in an attempt to predict the stock prices. These attempts using the various ML and DL algorithms are based on the analysis of the historical data and do not factor the diverse elements affecting the stock market. The stock market prices and trends can be predicted more accurately and consistently only if all the factors affecting the company stocks such as news sentiment, socio-political developments etc. are taken into account. Besides being unable to factor all the disparate factors influencing the market, the various ML and DL algorithms have their own individual limitations and drawbacks. In the domain of Machine Learning, the algorithms such as the RF Classifier, KNN or Decision Trees all fail to function as efficient regressors and deliver poor accuracy in predicting the stock prices. The most popular SVM Regressor is a more accurate and consistent ML algorithm but has its limitations. The results it delivers are dependent on how its parameters are tuned and optimized and the selection of an appropriate kernel poses a big problem in time-series problems. In the domain of Deep Learning, RNN and ANN networks are a popular choice when it comes to predicting the stock prices. These two have their limitations and drawbacks such as the Vanishing Gradient problem of RNN.

This work proposes two models that have a similar approach. The first model is a Hybrid ML model which integrates an SVM regressor, KNN classifier, an RF classifier and a Linear Regressor using a Majority Voting algorithm. The other model is a Hybrid LSTM model optimized by a backpropagation network. Both models are continuously trained using Reinforcement Learning.

The models are given rewards according to the accuracy of their predictions and results. By employing Reinforcement Learning, the models continue to learn as they function and their accuracy and consistency increase continuously. Both models use the various technical indicators like RSI, Bollinger Bands, OHLC, EMA etc. as additional parameters. A Sentiment Analyzer is a part of both the models, that factors the news sentiment affecting the stock market for more accurate results.

## II. RELATED WORK

Various models, systems and algorithms have been utilized in an attempt to predict the stock market trends. The SVM has been one of the most popular algorithms employed in the field of ML, [1] the Support Vector Regressor (SVR) is profusely utilized to predict the market trends as it possesses features like noise-tolerance, tuning of hyperparameters and gives a pretty decent accuracy. KNN is among the other ML algorithms used, [2 ANNs have been used conventionally, to predict share price trends but have proved to be inefficient due to their limitations like over-fitting of data over time. The Recurrent Neural Network (RNN) became very a popular choice for market prediction, [3] KNN has also been used in an attempt to predict share market prices to some extent but with poor accuracy. In the domain of DL, ANN was among the first DL algorithms used to predict and analyze the share market [4] Recurrent Neural Networks have been widely employed to solve the share price prediction problem and gives good accuracy which varies according to the epochs used to train the model. The RNN has been found to be inaccurate in predicting the market prices due to its limitations like the Vanishing Gradient problem. Among other approaches taken [5] to predict share market prices, a common one is based on the Sentiment Determination of the financial and socio-political news impacting to the market Various comparative studies and researches have been conducted to compare and contrast the disparate ML and DL algorithms and approaches.[6] The traditional and financial approach for decades has been to use the financial Technical Indicators to predict the market prices.

### A. Support Vector Regressor (SVR)

The SVR is among the supervised ML algorithms which is employed both for prediction and classification problems. In the SVM algorithm, each entry of data is taken as a point in a multi-dimensional space (which is dependent on the number of features taken ‘n’) and the value of the point is its coordinates the space. SVR employs kernel functions, which are mathematical functions, for classification and regression. The Kernel takes and transforms input data into a suitable classified form. Three different Kernel Functions: RBF, Polynomial, Linear and Sigmoid.

### B. KNN Classifier Algorithm

The K-Nearest Neighbor is a relatively simpler technique that too can be used to solve the share price prediction problem. It is much more efficient and powerful when compared to the conventional Supervised Learning algorithms like Linear Regression, RF Classifier, and other such ML algorithms. It takes a data sample as input with a defined class and an objective dataset of an undefined class for classification. The algorithm computes which defined classes are closest and most similar to the unknown target class (nearest neighbors) attempting to categorize the undefined class. If the ‘K’ value is equal to 1, then we will categorize the target unknown class using the single nearest neighbor class.

### C. Majority Voting Algorithm

Majority Voting technique integrates different regressors and classifiers to give higher accuracy and better results. Firstly, different classifiers or regressors are created for varying subsets of the input data. Each created regressor makes and gives an output prediction. The results of every model are stored in a matrix or an n-dimensional array. These output predictions of the regressors is defined as a single vote. The multi-dimensional array is then iterated upon and the instances of every output prediction are counted and documented. The resultant prediction is the prediction whose instances have votes greater than half of the total.

### D. LSTM (Long-Short Term Memory)

The LSTM networks happen to be a variant of the RNN systems that help in tackling the Vanishing Gradient problem possessed by the Recurrent Neural systems. As in Recurrent Neural systems, here too, there are time steps but an additional feature “memory” is included for all time steps. Cells are a fundamental part of the LSTM network which are basically memory blocks. The former cell or the previous cell gives two states as input to the latter cell viz the hidden state along with the cell state. Three gates are employed for processing large data viz Input Gate, Forget Gate and, the Output Gate. A forget gate removes the unnecessary data from the cell state. It employs the sigmoid function to append new data and also defines a vector containing all the possible values that may be appended to the current cell state.

## III. DEVELOPED MODEL

In this work, models are developed for share price prediction viz a Hybrid ML Model along with a backpropagation optimized LSTM model. The models utilize the historical stock market data, the technical indicators that are the additional parameters, as well as the financial, social and political news sentiment impacting the share market. The models are trained continuously using Reinforcement Learning to increase the consistency and accuracy. The financial, political and social news is collated from various trusted sources and platforms like The Economist, The Financial Times, BBC, CNN, Twitter and many more. This collated data processed using NLP techniques.

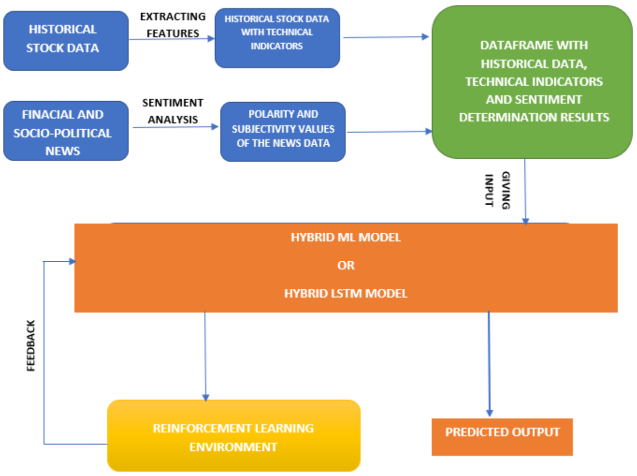
HISTORICAL STOCK DATA
EXTRACTING FEATURES
HISTORICAL STOCK DATA WITH TECHNICAL INDICATORS
FINACIAL AND SOCIO-POLITICAL NEWS
SENTIMENT ANALYSIS
POLARITY AND SUBJECTIVITY VALUES OF THE NEWS DATA
DATAFRAME WITH HISTORICAL DATA, TECHNICAL INDICATORS AND SENTIMENT DETERMINATION RESULTS
GIVING INPUT
HYBRID ML MODEL
OR
HYBRID LSTM MODEL
FEEDBACK
REINFORCEMENT LEARNING ENVIRONMENT
PREDICTED OUTPUT

**Figure 1: Developed "Intellifin" Model**

### A. Collation of Data

The ***Historical Data for stocks*** is collated using multiple APIs viz YFinance and NSE APIs. Both the APIs are employed to gather past share market statistics.

***The Real-Time Data*** is collated using the AlphaVantage API for the various global stock exchanges. The real-time information helps predict the next-minute stock prices and assist intraday trading.

***The Financial, Political and Social Data for Sentiment Analysis*** is collated from various trusted sources and platforms like The Economist, The Financial Times, BBC, CNN, Twitter and many more using the News API and the Tweepy Twitter API.

**Table 1: Input Dataset for Apple Company STOCK (2019)**

| *DATE* | *HIGH* | *LOW* | *OPEN* | *CLOSE* | *VOL* |
|---|---|---|---|---|---|
| 12-16 | 277.00 | 280.79 | 276.98 | 279.86 | 32046500 |
| 12-17 | 279.57 | 281.77 | 278.80 | 280.41 | 28539600 |
| 12-18 | 279.80 | 281.90 | 279.12 | 279.74 | 29007100 |
| 12-19 | 279.50 | 281.18 | 278.95 | 280.02 | 24592300 |

### B. Data Preprocessing

The financial, social and political news data for Sentiment Determination that is collated from the various news and social media platforms is processed using NLP. The collated data in a textual format is split by individual words. These segregated words are compared with a collection of words that include a defined dictionary of lexicons to classify them into three broad categories viz negative, positive and neutral. The special symbols and other symbols such as emoticons are mined along with the words from the textual input to assist classification. The Naïve Bayes Classifier is used to categorize the data into the three defined categories. The NLTK library is then employed to analyze and determine the Subjectivity and Polarity values of the sample input textual data.

The historical market data collated is processed to identify and find any incomplete data entries or missing or null values. The data is then visualized by plotting it in s graphical form. Visualizing the data provides the required insight about the data and an absolute idea about the input data. These incomplete entries or null entries are either or techniques such as Regression are employed to predict these values. This processed data is stored in a separate dataframe.

The subjectivity and polarity results from the financial and socio-political news data is appended as a new row to the input dataframe that contains the historical market data. This makes these results an additional feature employed to train both the developed models.

**Table 2: Apple Company Data frame After Sentiment Determination**

| *DATE* | *HIGH* | *LOW* | *OPEN* | *CLOSE* | *VOL* | *POL.* |
|---|---|---|---|---|---|---|
| 12-16 | 277.00 | 280.79 | 276.98 | 279.86 | 32046500 | 0 |
| 12-17 | 279.57 | 281.77 | 278.80 | 280.41 | 28539600 | 0.0166 |
| 12-18 | 279.80 | 281.90 | 279.12 | 279.74 | 29007100 | 0 |
| 12-19 | 279.50 | 281.18 | 278.95 | 280.02 | 24592300 | -0.184 |

### C. Feature Extraction

To increase the accuracy of prediction, various financial Technical Indicators are used as additional parameters that are given as input along with historical and financial sentiment data. Technical indicators scientific calculations performed on the market data factors like volumes traded or OHLC prices which are used by financial market analysts. By observing previous market data, financial analysts use these technical indicators to predict share prices. The technical indicators used are:

1. Stochastic Oscillator
2. Bollinger Bands
3. Exponential Moving Average (EMA)
4. Money Flow Index (MFI)
5. Relative Strength Index (RSI)
6. Moving Average Convergence Divergence (MACD)

These indicators are computed using the historical stock price data and are concatenated to form a new dataframe consisting of all these indicators which is then given as input to developed models.

**Table 3: Sample Data frame after Feature Extraction**

| *DATE* | *HIGH* | *LOW* | *OPEN* | *CLOSE* | *VOL* | *POL.* | *RSI* | *EMA* | *MFI* |
|---|---|---|---|---|---|---|---|---|---|
| 16-Dec | 277 | 280.79 | 276.98 | 279.86 | 32046500 | 0 | 40 | 16 | 45 |
| 17-Dec | 279.57 | 281.77 | 278.8 | 280.41 | 28539600 | 0.0166 | 32 | 14 | 42.3 |
| 18-Dec | 279.8 | 281.9 | 279.12 | 279.74 | 29007100 | 0 | 28 | 14 | 35 |
| 19-Dec | 279.5 | 281.18 | 278.95 | 280.02 | 24592300 | -0.184 | 55 | 17 | 42 |

## IV. IMPLEMENTATION METHODOLOGY

The two models are implemented using the following methodology:

- Sentiment Determination
- Technical Indicators Extraction
- Implementing the Hybrid ML Model
- Implementing the Hybrid LSTM Model
- Creating the Reinforcement Learning Environment

### A. Implementing Sentiment Analysis

The sentiment determination is implemented using the NLTK Library for on the data gathered using the News API as follows:

I. Register for News API
II. Initiate the Tweepy API
III. Set the sources on the News API
IV. Include the required libraries and files
V. Collate the financial and socio-political news for the target company from the specified sources using the News API
VI. Using NLTK library, compute the polarity and subjectivity of each data entry
VII. The results are collated into a dataframe and in an offline file

First, the News API key is acquired to collate data. The gathered data is then processed using the NLTK library. Firstly, the Naïve Bayes Classifier is used to classify the data as positive, neutral and negative. Then the TextBlob library is used to analyse the polarity and subjectivity of each all entries. These results are collated as a dataframe for processing. This dataframe is then concatenated with the data-frame holding the historical stock market data to form a single data-frame.

### B. Implementing the Hybrid Ml Model

The Hybrid ML model integrates an SVM regressor, KNN classifier, an RF classifier and a Linear Regressor using a Voting Ensembler as follows:

I. Include the required libraries
II. Initiate the YFinance and the AlphaVantage APIs
III. Import the required historical data using the YFinance API
IV. Initiate the SVM Regressor
V. Initiate the KNN Classifier
VI. Initiate the RF Classifier
VII. Create a Voting Ensemble Classifier
VIII. The created ML classifiers and regressors are passed to the Ensemble Classifier as input
IX. Obtain the combined output from the Ensemble classifier
X. Compute the confidence score

The required libraries and files are imported. The YFinance and AlphaVantage APIs are utilized to collate historical data as well as the intraday data. The required classifiers viz SVM, RF, KNN are created on the different samples of the same input data. The Majority Voting Classifier is used to combine all the created classifiers and assign weights to each.

### C. Implementing the Hybrid LSTM Model

The LSTM Algorithm is implemented in Python using Keras and TensorFlow backend in the following steps:

I. Include the required packages
II. Give as input, the developed input data sample
III. Transform the dataframe into an array
IV. Divide the data into two parts viz 80% data will be used for training the developed model and 20% data will be used for testing the model
V. The different LSTM layers are created to form the network model
VI. Initialize the created layers and compile all the layers into a single network
VII. Compute the confidence score

This model starts by importing the created input sample data-frame along with the past market statistics and Sentiment Determination resultant values. These data-frames are collated in a multi-dimensional array. This data is then segregated into two parts viz 80% data that will be used for training the developed model called the training set and 20% data that will be used for testing the model called the test set. The different layers of the model are defined.

### D. Creating the Reinforcement Learning Environment

The Reinforcement Learning is implemented using the Q Learning Algorithm in the OpenAI Gym Environment in the following steps:

I. Register on the OpenAI Gym platform
II. Install all Gym dependencies
III. Define the Q-Learning table
IV. Define the type and shape of the action space
V. Define the observation space that will contain the data to be observed
VI. The Open, High, Low, Close and Adj, Volume are the parameters to be observed in the observation space

The Reinforcement Learning is implemented by using the Q-Learning Algorithm in the OpenAI Gym environment. The gym package is used for the purpose. The action space is defined which defines the Q-Learning Table. The observation space is defined next to declare the Open, High, Low, Close and Adj. Volume as the parameters to be observed by the agent.

## V. RESULT AND ANALYSIS

The above models were executed upon the past market statistics and data of companies. The sentiment determination results of the news data related to the stock market was also combined with historical data to give the combined dataframe. The additional extracted technical indicator parameters are also computed and integrated in the dataframe to give the final input data sample. The models are trained continuously using the Q-Learning Algorithm.

### A. Hybrid Ml Model Result

The Hybrid ML model proved to be very effective and efficient in predicting the next-minute share prices and hence serves as an excellent tool for intraday trading. It was efficient in predicting the next day stock prices with a varying accuracy ranging between 92-97%. It proves to be extremely consistent and efficient in predicting market prices up to a week with accuracy in the range of 91-97%. It has limitations and drawbacks when it comes to predicting long term market trends and the accuracy falls when attempting to predict for the next month or longer

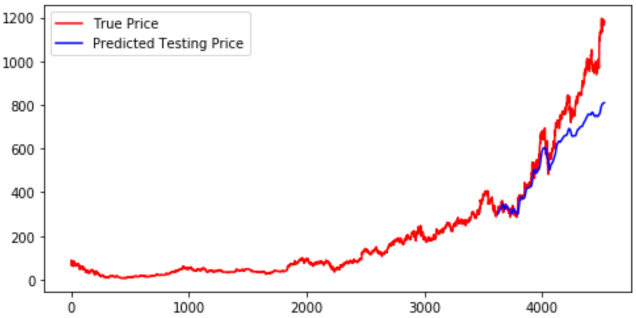

**Figure 2: Predicted Share Trend for Amazon – Hybrid Ml Model**

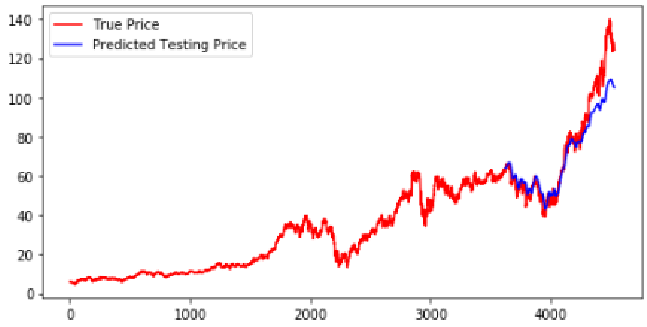

**Figure 3: Predicted Share Trend For Adobe– Hybrid Ml Model**

### B. Hybrid LSTM Model Result

The Hybrid LSTM model predicted next-day stock prices with an outstanding accuracy ranging between 94-99%. It proves to be extremely accurate and efficient in predicting the next-day prices for up to a week with the accuracy not falling below 95%. It also overcomes the limitation of the Hybrid ML model and proves to be efficient in predicting long-term market trends, and correctly predicts market trends for up to 4 weeks. It hence is an outstanding tool for aiding long-term investment in the share market. It is not very effective while predicting the next-minute share prices and hence not a suitable tool for intraday trading.

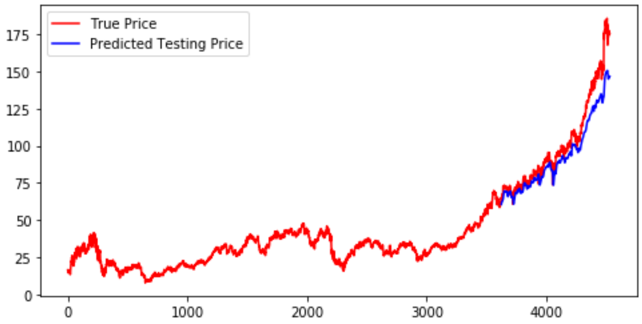

**Figure 4: Predicted Share Trend for Amazon – Hybrid LSTM Model**

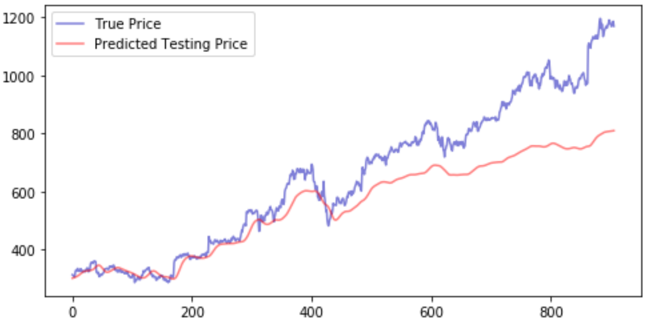

**Figure 5: Predicted Share Trend for Reliance – Hybrid LSTM Model**

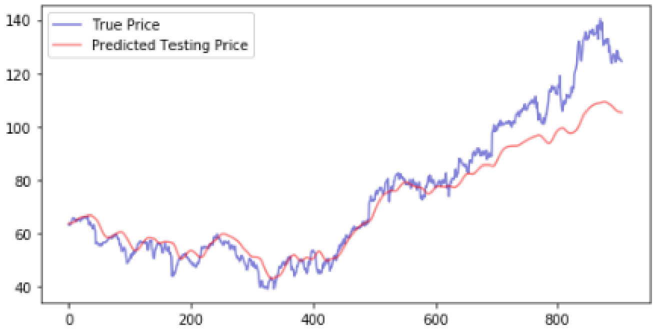

**Figure 6: Predicted Share Trend for Tata Steel– Hybrid LSTM Model**

## VI. CONCLUSION AND FUTURE WORK

The Stock market is an erratic system wherein copious data is produced at frequent intervals and fluctuates rapidly due to several disparate factors. Various models, systems and algorithms have been utilized in an attempt to predict the stock market trends. Many of these belong to the domain of ML and DL. Predicting the market trends and direction by analyzing the financial and social news that impacts the market using NLP and Sentiment Analysis is another very widely used approach. The traditional and financial approach for decades has been to use the financial Technical Indicators to predict the market prices.

This work develops two models that function on the novel developed "IntelliFin" algorithm that takes in three input datasets viz Historical market statistics and data, The financial-socio political news, rumors and happenings impacting the market and the Financial Technical Indicators that have been conventionally used since decades by the market analysts to predict the market prices. It then processes these input datasets and concatenates them into one sample input dataset and gives a prediction of market prices as output. The system also employs Reinforcement learning algorithms to continuously train the models as they predict and function and make them more accurate and efficient.

Both developed models have their individual plus points and advantages and prove to be very efficient. They both complement each other as they overcome the drawbacks of one another and hence can serve as an extremely efficient and helpful tool for financial investments, financial analysis and market prediction.

## REFERENCES

1. Jui-Sheng Chou and Thi-Kha Nguyen, “Forward Forecast of Stock Price Using Sliding-window Metaheuristic-optimized Machine Learning Regression”, DOI 10.1109/TII.2018.2794389, IEEE Transactions on Industrial Informatics
2. Min Wen, Ping Li, Lingfei Zhang, and Yan Chen, “Stock Market Trend Prediction Using High-Order Information of Time Series”, date of publication February 26, 2019, date of current version March 18, 2019. 10.1109/ACCESS.2019.2901842
3. Yongsheng Ding, Lijun Cheng, Witold Pedrycz, and Kuangrong Hao, “Global Nonlinear Kernel Prediction for Large Data Set With a Particle Swarm-Optimized Interval Support Vector Regression”, Ieee Transactions On Neural Networks And Learning Systems, Vol. 26, No. 10, October 2015
4. L. Minh Dang, Abolghasem Sadeghi-Niaraki, Huy D. Huynh, Kyungbok Min And Hyeonjoon Moon,” Deep Learning Approach for Short-Term Stock Trends Prediction based on Two-stream Gated Recurrent Unit Network”, DOI 10.1109/ACCESS.2018.2868970, IEEE Access
5. Lei Shi, Zhiyang Teng, Le Wang, Yue Zhang, and Alexander Binder, “DeepClue: Visual Interpretation of Text-based Deep Stock Prediction”, DOI 10.1109/TKDE.2018.2854193, IEEE Transactions on Knowledge and Data Engineering
6. Guang Liu And Xiaojie Wang, “A Numerical-based Attention Method for Stock Market Prediction with Dual Information”, 10.1109/ACCESS.2018.2886367, IEEE Access
7. Rashmi Sutkatti, Dr. D. A. Torse, “Stock Market Forecasting Techniques: A Survey”, Volume: 06 Issue: 05 | May 2019, International Research Journal of Engineering and Technology (IRJET)
8. Divit Karmaini, Ruman Kazi, Ameya Nambisan, Aastha Shash, Vijaya Kamble, “Comparison of Predictive Algorithms: Backpropagation, SVM, LSTM and Kalman Filter for Stock Market”, 10.1109/AICAI.2019.8701258, IEEE
9. Priyamvada, Rajesh Wadhvani, "Review on various models for time series forecasting", Inventive Computing and Informatics (ICICI) International Conference on, pp. 405-410, 2017.
10. Aparna Nayak, M. M. Manohara Pai∗ and Radhika M. Pai, “Prediction Models for Indian Stock Market”

## AUTHORS PROFILE

**Shashank Singh** is currently pursuing B.Tech. Degree in Computer Science and Engineering from SRM Institute of Science and Technology, Chennai. He has completed his Secondary Education from Amity International School, Mayur Vihar, Delhi. He has published a paper in the domain of cognitive learning titled “*Predicting Stock Market Trends using Hybrid SVM Model and LSTM with Sentiment Determination using Natural Language Processing*” in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019. He has completed several projects in the field of Machine Learning and Deep Learning.

**Maaz Ahmad** is studying Computer Science and Engineering in SRM Institute of Science and Technology, Chennai. He has completed his Secondary Education from D.G. Sr. SEC. +2 School Dumariya, Gopalganj, Bihar.

**Aditya Bhattacharya** has completed his secondary education from Sri Chaitanya Institute, Vishakhapatnam, Andhra Pradesh. He is completing his Computer Science B.Tech. Degree from SRM Institute of Science and Technology, Chennai

**D. Prabhu** has completed his M.E in Computer And Communication from Sri Sai Ram Engineering college, Chennai and B.E in Computer Science and Engineering from Annamalai University, Chidambaram and is currently pursuing his P.h.D. in SRM Institute of Science and Technology, Chennai and working as Assistant Professor in Computer Science and Engineering at SRM Institute of Science and Technology, Chennai. He has authored a Paper on An Implementation of Secure Authorized Elimination of Duplicate Copies of Files Using Security Enabled Approach Proceedings of IJRECS @ Aug – Sep 2016, V-6, I-1ISSN-2321-5485 (Online) ISSN-2321-5784 (Print)