Journal of Informatics Education and Research
ISSN: 1526-4726
Vol 5 Issue 1 (2025)

A COMPREHENSIVE HYBRID MACHINE LEARNING MODEL FOR STOCK PRICE PREDICTION

Roopa Devi E M
Department of Information Technology,
Kongu Engineering College Perundurai, Erode, Tamil Nadu
roopadevi.it@kongu.edu
Elakya K
Department of Information Technology,
Kongu Engineering College Perundurai, Erode, Tamil Nadu
elakyak.21it@kongu.edu
Diyaneish M S
Department of Information Technology,
Kongu Engineering College Perundurai, Erode, Tamil Nadu
diyaneishms.21it@kongu.edu
Hamshavarthini T
Department of Information Technology,
Kongu Engineering College Perundurai, Erode, Tamil Nadu
hamshavarthinit.21it@kongu.edu

ABSTRACT:
With increasing numbers of investors into the stock market, knowing how to reduce the risk associated with it has become a matter of high importance. Determinant predictive models for stock price movements using machine learning methods have been developed in this regard. This article uses data obtained from the National Stock Exchange of India to create a model which will predict the direction for stock price movement. The model should seek and make accurate and reliable predictions on the basis of mutual relationships between various market indicators. This all is based on identifying factors that have a significant impact on stocks and their futures as well as the better understanding of market dynamics that would lead to informed decisions by investors. The last part of this model is to explain how independent variables interact in order to ultimately influence future trends of the market which helps improve financial forecasting.

Keywords: Time series forecasting, Market data analysis, Logistic regression analysis, Ensemble learning.

I.INTRODUCTION

The stock market is one of the most crucial factors in the global economy, being an investment tool, a generator of wealth, and a stimulant to economic growth. Due to the dynamic nature of this sector, with its interrelated variables, including corporate performance, investor sentiment, and macroeconomic conditions, predicting the movement of stock prices becomes complex yet very important. Accurate stock price prediction can provide valuable insights to investors and analysts for mitigating risks and maximizing returns. This makes stock market volatility and nonlinearity an important challenge to the traditional methods of forecasting, which cannot always capture intricate relationships involving several influencing factors.

As a result, the ability to analyze massive datasets, uncover hidden patterns, and adapt to changing market conditions makes machine learning an enormously powerful tool in addressing these challenges. Most importantly, machine learning algorithms do model complex and nonlinear relationships; therefore, the sophistication of accuracy and reliability applied in the stock price prediction improves. Historical market data and advanced computational techniques create room for improving decision-making by assuming market uncertainty.

II. LITERATURE SURVEY

Gourav Kumar et al. In this system [1] have suggested that the use of two-way telecommunications systems is sufficient both to support and inhibit " The economic and social structure of a nation is heavily affected by its stock markets. The noise, volatility complexity nonlinearity dynamism and chaos of stock market time series make it difficult for investors analysts and researchers to forecast the stock market. Stock market forecasting is a crucial area of research in finance due to the high level risk associated with investing.
The great majority of the danger can be reduced with the introduction of computationally complicated technologies. This essay reviews recent studies on artificial intelligence techniques used for stock market forecasting. This article collects and examines chosen articles around the following major themes: (1) stock market dataset analysis and dataset, (2) input variables; (3) pre-processing approaches; (4) feature selection forecast models and (6) performance measures for model assessment. This study provides academics and financial analysts with a systematic approach to providing intelligent stock market forecasting tools. This research suggests enhancements to existing approaches.

Vinay Anand Tikkiwal et al. [2] proposed this system. As requirements for clean fuels such as hydrogen and gas rise solar energy is gaining popularity. However, Solar radiation is System performance is directly impacted by solar energy's stochastic and intermittent nature. With the sensitivity of solar radiation it is likely that the new solar cells will be better able to see exactly how much radiation is generated. This study outlines an innovative method for using LSTM networks to estimate day-ahead global horizontal radiation. Historical hourly data on weather parameters like temperature pressure, wind speed relative humidity dew point and wind direction were among the inputs. In our suggested model the LSTM and the enhanced Cuckoo Search approach are both included. Two statistical metrics—RMSE and MAE—have been used to assess the model's performance. Because of its low parameter values the suggested model performs better than existing models for day-ahead forecasting of global horizontal radiation as predicted.

Deep learning is being used to predict global solar radiation for dry zones in many steps. Deeksha Chandola et al. [3] have recommended in this system. Solar irradiation is variable and is intermittent in nature. It is therefore important to consider this unpredictability in order to maximize the use of solar energy. For solar energy-based systems to be designed and operated as efficiently as possible, solar radiation forecasting is helpful.

This study explores how psycholinguistic aspects of financial news can be used to predict movements in the Indian stock market. As noted by B. Shravan Kumar[4] and colleagues, using news content for financial forecasting is an emerging area of interest. In our research, we utilize psycholinguistic metrics derived from news articles—specifically, tools like LIWC and TAALES—as predictor variables to develop hybrid intelligence models for stock market predictions.

We applied various advanced techniques for these predictions. Our analysis focused on data from 12 companies listed on the Bombay Stock Exchange. To extract psycholinguistic features from the news articles, we implemented several feature selection methods, including chi-squared tests and maximum relevance with minimum redundancy (MRMR).After conducting a thorough evaluation using the Diebold-Mariano test, our findings indicated that GMDH and GRNN proved to be the most effective strategies, particularly when assessed by their MAPE and NRMSE values.

A New Approach to Predicting Market Trends Using Twitter and Financial News: A Study on the Istanbul Stock Exchange (BIST 100) ZEYNEP HİLAL KILIMC and his collaborators have

proposed an exciting system that can predict stock movements, exchange rates, and market trends—a hot topic for analysts, investors, and researchers. This study is about using word embedding and deep learning methods to forecast the market direction of the Istanbul Stock Exchange (BIST 100) by looking at nine famous banking stocks. In the past, English-language news was the most available data on market trends. To prove this, we got information from four well-known Turkish news sources. This content includes technical analysis from Bigpara, user comments from both Twitter and Mynet Finance, and stock-related headlines from the Public Disclosure Platform (KAP). Our experiments demonstrate that the use of deep learning combined with wordembedding algorithms would be a promising strategy in predicting the trends of the BIST 100.

III.PROPOSED METHODOLOGY

This method proposes within the machine learning algorithms predicting future stock prices and changes in stock rates using a large dataset from India' The system uses the relationships found in financial market data to produce accurate and reliable projections that help investors make well-informed decisions. This method considers the relationship between various financial indicators and how those relationships affect future market values.

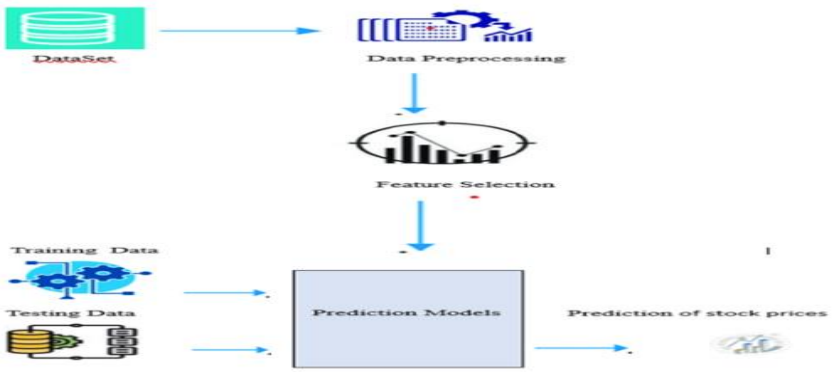

Fig.1.Prediction of stock price

3.1. Dataset Preparation and Loading

The system starts with importing the dataset in a standardized format, for example, CSV files to ensure compatibility with machine learning workflows.The raw dataset contains the necessary stock attributes, which include Date, Open, High, Low, and Close prices, along with the associated temporal features.The Date column, rather than being discarded, is transformed into new features like day, month, year, or even day of the week to capture seasonal trends and temporal effects that influence stock price movements.

3.2. Data Preprocessing

This process of cleaning and making it in a co-operative way suitable to model for final prediction forms preprocessing that is crucial within the pipeline for machine learning. Handling missing values will be the foremost step for cleanup. Without being addressed, the trained model poses many inconsistencies and inaccuracies with missing values within its core. Various imputation techniques - filling missing values with mean, median, mode, etc.; or removal of incomplete records depend on the character of missing data. These avoid unwanted biases while upholding the consistency of the dataset.

Based on the dataset developed, derived metrics are used in order to add more predictiveness in the data set. For example, percentage changes in Close prices are computed. This metric measures the changes day-to-day stock price movement following some temporal pattern and helps develop a target variable to build the classifier for this task. The target variable is encoded as a binary class: 1 for price increase and 0 for price decrease or no change. This binary representation of the target variable simplifies the prediction task and fits the objectives of the classification model.

Then, data scaling techniques such as Min-Max normalization or Z-score standardization are applied to prepare features for effective learning. Here, Min-Max normalization scales all feature values within a fixed range, usually [0, 1], thereby ensuring that no single feature dominates because of differences in magnitude. However, the feature Z-score standardizes in such a way that mean values equal to zero and the standard deviations are equal to one. That way, especially sensitive algorithms related to scaling are useful to deal with support vector machines as well as logistic regression.

3.3. Feature Selection

Important indicators such as Open, High, Low, Close, and time features from Date are kept as input variables since they have been proven to be related to the movement of stock prices. Redundant or irrelevant features are removed to make the model more straightforward and increase the efficiency of computation. Statistical methods, such as correlation analysis, are used in order to ensure that the features selected are also predictive relevant.

3.4. Machine Learning Models

This module separates the data into training and testing sets. This involves training machine learning models, such as SVM and Logistic Regression, on training data, after which their performance is evaluated on testing data in order to see if they do generalize well to new data.

3.4.1 Decision Tree:

The final classification outcome is represented by the leaf nodes of a decision tree, an ordered model used in supervised learning. Potential decision paths are indicated by the branches, and the many data aspects are reflected in the interior nodes. When evaluating the degree of unpredictability or impurity in a dataset, entropy is employed.

$$E(S) = -P_{(+)}logp_{(+)} - P_{(-)}logp_{(-)} \quad (1)$$

3.4.2 Random Forest:

It is quite well-liked since it performs well even with the default hyperparameters. It constructs several decision trees utilizing several dataset samples as opposed to depending just on one.

$$y = \frac{1}{N}\sum_{i=1}^{N} T_i\,(x) \qquad (2)$$

All of the trees' predictions are pooled, and a voting method is used to determine which forecast is the final output.

3.4.3 Support Vector Machine:

To separate data points from distinct classes as effectively as feasible for precise predictions, it finds the ideal hyperplane that optimizes the margin between them. The function that makes decisions is:

$$predicted\ output = sign(w.X + b) \quad (3)$$

3.4.4 K-Nearest Neighbors:

This approach for classification is instance-based. By taking into account the majority class among a data point's k-nearest neighbors—which are identified using a distance measure like Euclidean distance—it allocates a class to the data point.

3.4.5 Logistic Regression:

This statistical method models the link between multiple independent factors and a dependent variable by fitting a regression curve to best represent the data. A predicted probability of a given classification is actually the output from applying the logistic function to some linear combination of the input variables.

$$predicted\ output = \frac{e^{(i_0+i_1X)}}{1+e^{(i_0+i_1X)}} \qquad (4)$$

3.4.6 Stacking Classifier:

Stacking is an ensemble learning technique where the outcome from more than one base model is combined in order to increase prediction accuracy.This approach makes use of several models such as Support Vector Classifier (SVC) and Logistic Regression for the purpose of making predictions. The final output is generated through training a more complex model known as meta-models using the predictions. Because different machine learning models have different biases, variances, and strengths, stacking aims to optimize performance. The goal is to improve the overall performance of the model ensemble by leveraging the benefits of different base models.

By combining SVM with Logistic Regression as a meta-model, one may obtain the advantages of both the models to get considerably improved predictive performances. For minimization of the total number of errors and achieving better generalization, the meta-model Logistic Regression appropriately combines the basis models' predictions.

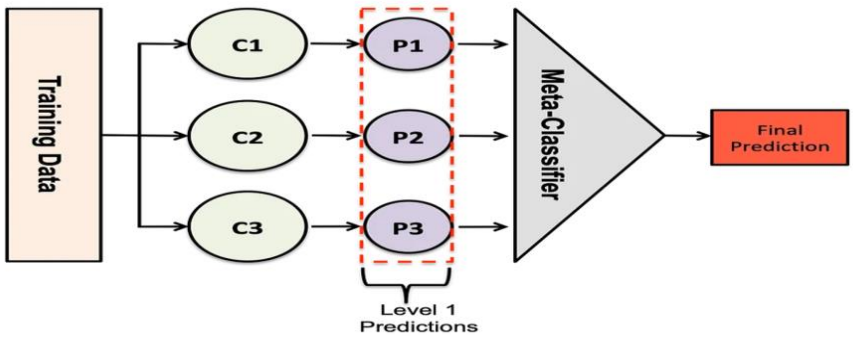

Fig.2.Stacking Classifier

IV. RESULTS AND ANALYSIS

The model results and the performance indicators will be graphically presented by this module.This includes making accuracy and MSE comparison graphs, the visual comparison of actual and projected stock values, and produce report summaries that give investors a summary of results and actionable insights for future investments.It's an improvement of the prediction in stock prices by applying methods from Support Vector Machine (SVM) and Logistic Regression along with stacking, since the algorithm classifies the data or continuous values to be predicted with a large margin, due to its capability to find the best hyperplane. The lower-dimensional input data will be transformed into a higher-dimensional feature space, further passed into the applied kernel function for the discovery of complex patterns.

Support Vector Machine Performance:

The confusion matrix clearly shows and indicates that this model's accuracy is 89.87%.

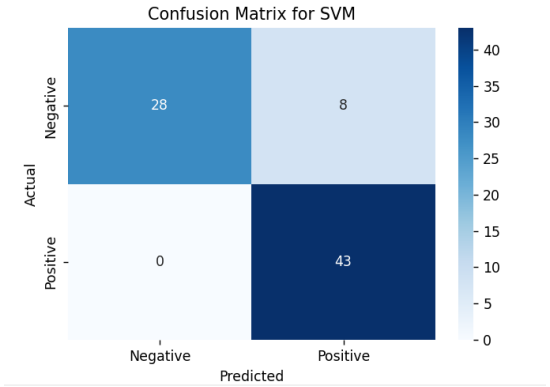

Fig.3.Support Vector Machine Performance

Logistic Regression Performance:

The confusion matrix clearly shows and indicates that this model's accuracy is 92.41%.

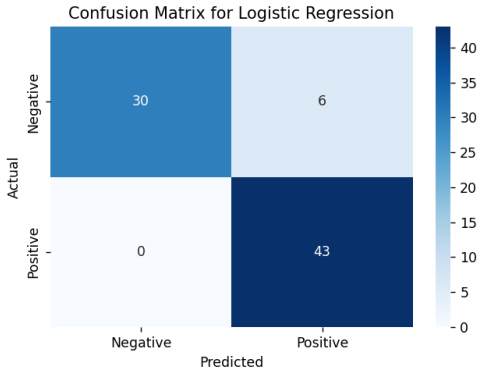

Fig.4.Logistic Regression Performance

The accuracy of the models are displayed in the below Fig.1.It shows that the Hybrid model using stacking has the highest accuracy among those models.

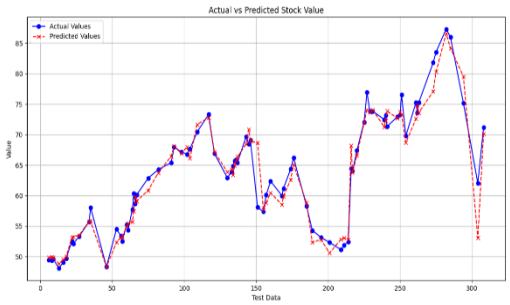

Fig.5. Comparison between actual and Predicted Stock Values

The outcome study showed that while stacking had an accuracy of 94.13%, the SVM and Logistic Regression models both had quite high accuracy, scoring 89.87% and 92.41%, respectively. These impressive accuracy rates demonstrate how well the models predicted changes in stock prices. However, the real vs. anticipated graphics demonstrate that although the models were largely accurate, there were still some discrepancies between the actual and expected stock prices. These graphics help identify certain scenarios where the models' predictions and actual outcomes

diverged, providing important insights into areas that require further research and refinement.

Table.1. Performance Analysis of various ML models

Machine Learning Models	Accuracy	Precision	Recall	F1-Score
Decision Tree	68.35	0.67	0.84	0.74
Random Forest	74.68	0.71	0.91	0.80
KNN	75.95	0.71	0.93	0.81
SVM	89.87	0.84	1.0	0.91
Logistic Regression	92.4	0.88	1.0	0.93
Stacking (SVM+ Logistic Regression)	94.13	0.89	1.0	0.94

V. CONCLUSION

Finally, utilizing SVM, Logistic Regression and stacking at the usage of these models for stock price prediction showed high accuracy rates. These results demonstrate that both models can correctly forecast changes in stock prices. However, the comparison of the actual and expected values reveals that there are still discrepancies, indicating areas that may benefit from further development. To improve the performance of the models, more feature engineering, data pre-processing, and parameter adjustment may be needed. All things considered, the project successfully offers a solid foundation for stock price prediction and offers informative data regarding the model's effectiveness; however, further adjustments may increase accuracy and better capture market dynamics.

VI.REFERENCES

[1] Kumar et al. (2020) conducted a survey on stock market forecasting using artificial intelligence. Archives of computational approaches in engineering, 1–33. https://doi.org/10.1007/s11831-020-09413-5.

[2] Tikkiwal VA, Vir Singh S, Gupta HO (2020). Day-ahead solar irradiance forecasting using a hybrid enhanced cuckoo search-lstm technique. In: 2020 2nd International Conference on Advances in Computing, Communication, and Networking (ICACCCN), pp. 84-88. Annals of Data Science (2023) 10(5): 1361-1378. https://doi.org/10.1109/ ICACCCN51052.2020.9362839.
[3] Chandola D, Gupta H, Tikkiwal VA, and Bohra MK (2020). Deep learning is used to anticipate global solar radiation for dry zones in several steps ahead of time. Procedia Computer Science 167:626–635. https://doi.org/10. 1016/j.proc.2020.03.329.

[4] Kumar BS, Ravi V, Miglani R. (2021). Predicting the Indian stock market based on psycholinguistic aspects of financial news. Ann Data Sci 8(3):517–558. https://doi.org/10.1007/s40745-020-00272-2

[5] Kilimci ZH, Duvar R. (2020). An effective word

embedding and deep learning-based model for forecasting the direction of the stock exchange market using Twitter and financial news sites: a case study of the Istanbul Stock Exchange (BIST 100). IEEE Access 8,188186–188198.

[6] Kumar, D., Sarangi, P. K., & Verma, R. (2022). A systematic review of stock market prediction using machine learning and statistical techniques. *Materials Today: Proceedings*, *49*, 3187-3191.

[7] Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. *Expert Systems with Applications*, *197*, 116659.

[8] Wu, Y., Fu, Z., Liu, X., & Bing, Y. (2023). A hybrid stock market prediction model based on GNG and reinforcement learning. *Expert Systems with Applications*, *228*, 120474.

[9] Jiang, M., Jia, L., Chen, Z., & Chen, W. (2022). The two-stage machine learning ensemble models for stock price prediction by combining mode decomposition, extreme learning machine and improved harmony search algorithm. *Annals of Operations Research*, 1-33.

[10]Mintarya, L. N., Halim, J. N., Angie, C., Achmad, S., & Kurniawan, A. (2023). Machine learning approaches in stock market prediction: A systematic literature review. *Procedia Computer Science*, *216*.