Predicting Stock Direction: A Comparative Study of Classifiers Using Technical Indicators and Machine Learning (December 2025)

FARAZ HEYDAR[1], HESAMEDDIN BITARAFAN RAJABI[2], AND SAAD BIN AHMED.[3]

[1]Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada (e-mail: fheydar3@lakeheadu.ca).
[2]Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada (e-mail: hbitaraf@lakeheadu.ca).
[3]Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada (e-mail: sbinahm@lakeheadu.ca).

ABSTRACT This project investigates the predictability of the Reliance Industries Limited stock price direction using a suite of machine learning algorithms. Challenging the Efficient Market Hypothesis, this study employs a rigorous feature engineering pipeline derived from thirteen technical indicators, including Moving Average Convergence Divergence (MACD), Triple Exponential Moving Average (T3), and Volatility Stop (VTS). To mitigate the curse of dimensionality and multicollinearity inherent in financial time-series data, Principal Component Analysis (PCA) is applied, reducing the feature space from 19 dimensions to 4 principal components while retaining 95% of the variance. We evaluate five distinct classifiers—Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest, and Multi-layer Perceptron (MLP)—and propose a Soft Voting Ensemble method to improve predictive stability. Experimental results covering the period 1994-2025 demonstrate that while the stock market exhibits significant stochastic behavior, the proposed Ensemble method achieves an Area Under the Curve (AUC) of 0.5256 and Accuracy of 52.45%, outperforming the baseline. Furthermore, feature importance analysis reveals that Trading Volume and Momentum indicators (CCI, RSI) possess higher predictive power than traditional trend-following metrics for daily forecasting.

INDEX TERMS Machine Learning, Stock Prediction, Technical Indicators, Ensemble Learning, PCA.

I. INTRODUCTION

THE prediction of stock market trends remains one of the most challenging tasks in financial computing due to the market's dynamic, non-linear, and stochastic nature [1]. Financial time-series data is inherently noisy and influenced by a multitude of complex factors, ranging from macroeconomic policy to investor psychology, making it difficult to model using linear assumptions [2]. Consequently, accurate forecasting requires robust computational frameworks capable of identifying hidden patterns within this volatility.

Historically, researchers relied on traditional statistical models such as Autoregressive Integrated Moving Average (ARIMA) and Generalized Autoregressive Conditional Heteroskedasticity (GARCH). While effective for stationary data, these models often struggle to capture the non-linear dependencies and abrupt structural breaks characteristic of modern financial markets [2]. This limitation has driven a paradigm shift toward Machine Learning (ML) and Deep Learning (DL) techniques, which possess superior capabilities in feature extraction and pattern recognition [1].

The primary objective of this study is to construct a predictive model capable of forecasting the daily directional movement (Up/Down) of Reliance Industries Limited stock prices. Unlike traditional econometric approaches, this project leverages a diverse suite of Machine Learning algorithms—including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest—to extract non-linear patterns from historical data. Furthermore, we address the common challenges of overfitting and multicollinearity by implementing a rigorous pipeline that includes Principal Component Analysis (PCA) for dimensionality reduction and a Soft Voting Ensemble for improved

generalization.

We focus on the Reliance Industries dataset covering the period from 1994 to 2025. The project implements a comprehensive methodology including:

1) Data Cleaning: Handling missing values using financial forward-fill methods to preserve temporal continuity.
2) Advanced Feature Engineering: Mathematical extraction of 13 distinct technical indicators (e.g., MACD, RSI, Bollinger Bands).
3) Dimensionality Reduction: Application of PCA to mitigate the "curse of dimensionality" by reducing the feature space while retaining 95% of the variance.
4) Ensemble Modeling: Integrating diverse classifiers to reduce variance and enhance predictive stability compared to individual models.

II. LITERATURE REVIEW

A. THEORETICAL FOUNDATIONS: EMH VS. PREDICTABILITY

The predictability of stock markets has long been a subject of intense academic debate. Fama introduced the Efficient Market Hypothesis (EMH), which asserts that asset prices fully reflect all available information, rendering it impossible to consistently outperform the market [1]. This view is often supported by the Random Walk Theory (RWT), which posits that price changes are serially independent and random [1]. However, empirical research has increasingly challenged these hypotheses, arguing that markets exhibit varying levels of efficiency (weak, semi-strong, and strong). Critics suggest that financial time series contain repeating patterns and trend determinism that can be exploited using advanced computational methods [3].

B. EVOLUTION FROM STATISTICAL TO MACHINE LEARNING MODELS

In the early stages of financial forecasting, scholars predominantly employed linear statistical models. For instance, researchers frequently utilized ARMA and GARCH models to forecast stock returns [2]. However, as financial markets evolved, the complexity and non-linearity of the data rendered these traditional econometric models inadequate.

Machine Learning algorithms have since emerged as a superior alternative due to their ability to process high-dimensional data and model complex, non-linear relationships. Wang et al. [1] discuss literature demonstrating that ML method such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) significantly outperform linear discriminant analysis in trend forecasting. Similarly, Fu and Zhang [2] noted that modern computational power allows for the processing of vast datasets, enabling models to learn intrinsic correlations that traditional methods miss.

C. ADVANCED CLASSIFIERS IN FINANCE

Recent literature has identified several key algorithms that are particularly effective for financial tasks:

- Support Vector Machines (SVM): SVMs are widely used for their ability to minimize structural risk. By using kernel functions (such as the Radial Basis Function), SVMs can map input data into high-dimensional feature spaces to find optimal hyperplanes for classification, making them robust against overfitting in smaller datasets [1].
- Ensemble Learning (Random Forest): Single classifiers often suffer from high variance. Ensemble methods, such as Random Forest, address this by constructing multiple decision trees on different subspaces of the data. Fu and Zhang [2] emphasize that ensemble approaches, including XGBoost and Random Forest, provide better generalization and stability than individual decision trees. This project builds on this insight by proposing a Voting Ensemble to aggregate predictions.
- Deep Learning (MLP & LSTM): While traditional ML remains powerful, Deep Learning architectures like Multi-layer Perceptrons (MLP) and Long Short-Term Memory (LSTM) networks have gained traction. LSTM networks, in particular, are designed to overcome the vanishing gradient problem, making them highly effective for capturing long-term dependencies in time-series data [2]. Although this study primarily utilizes MLP and traditional ML ensembles, the literature suggests that integrating these with deep learning architectures is a promising frontier for minimizing prediction error [1].

D. FEATURE ENGINEERING AND DIMENSIONALITY REDUCTION

The success of any ML model relies heavily on the quality of input features. Technical indicators—such as Moving Averages and Momentum oscillators—are standard proxies for market psychology [3]. However, using too many correlated indicators can lead to multicollinearity. Previous works demonstrate that dimensionality reduction techniques like Principal Component Analysis (PCA) are essential for filtering noise and improving model convergence speed without sacrificing significant information.

III. METHODOLOGY

A. DATA COLLECTION AND PREPROCESSING

The dataset comprises historical daily stock data for Reliance Industries from 1994 to 2025 obtained from Kaggle [1] offering over three decades of financial evolution from one of India's largest conglomerates. It includes raw market data and engineered features for time series modeling, volatility analysis, and signal generation. This dataset is structured for high-impact insights.

The raw dataset contained only five fundamental features:

- Date – Trading day in YYYY-MM-DD format
- Symbol – Stock ticker (e.g., RELIANCE)
- Open, High, Low, Close – Daily OHLC prices

[1] https://www.kaggle.com/datasets/jatinkalra17/reliance-30-years-of-market-data19942025

- PrevClose – Previous day's closing price
- Volume – Number of shares traded
- Turnover – Total traded value in
- VWAP – Volume Weighted Average Price
- Trades – Number of trades executed
- Daily Return – Daily percentage return
- MA 20 – 20-day moving average of closing price
- MA 50 – 50-day moving average of closing pric

Since the raw data did not contain a predictive label, Trades and Daily Return. we mathematically derived a binary target variable (y_t) to represent the daily directional movement.

$$y_t = \begin{cases} 1 & \text{if } Close_{t+1} > Close_t \text{ (Up)} \\ 0 & \text{otherwise (Down)} \end{cases} \quad (1)$$

To ensure the validity of our accuracy metrics, we analyzed the distribution of this derived target. Fig. 1 illustrates the class balance. The dataset exhibits a near-even split between 'Up' (51%) and 'Down' (49%) movements, confirming that the model will not be biased toward a majority class.

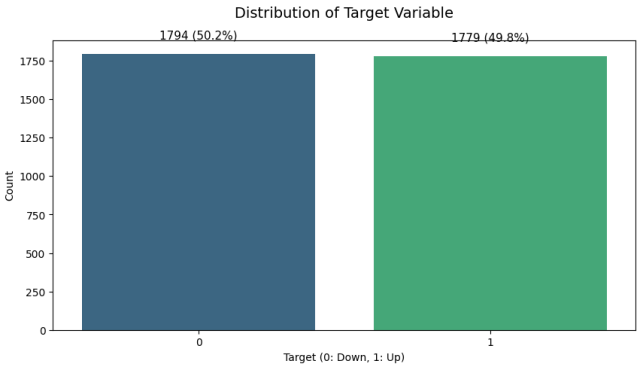

FIGURE 1. Target Class Distribution. The balanced nature of the derived target variable prevents class imbalance bias during training.

B. FEATURE IDENTIFICATION AND EXTRACTION

As the raw OHLCV data is insufficient for capturing complex market dynamics, we used engineered and comprehensive set of 13 technical indicators [3] along with OHLC and volume. These features were not present in the original dataset and were calculated to capture specific market dimensions:

1) Trend (SMA, EMA, T3)
2) Momentum (RSI, CCI, MACD)
3) Volatility (Bollinger Bands, VTS)
4) Volume (OBV)

C. EXPLORATORY DATA ANALYSIS (EDA)

Before modeling, we conducted extensive exploratory analysis to understand the data structure, including time-series visualization and correlation checks.

A Pearson correlation heatmap (Fig. 2) was generated to evaluate the relationships between our engineered features. The analysis revealed distinct clusters of high correlation (coefficients > 0.95), particularly between trend indicators like SMA, EMA, and the middle Bollinger Band. This redundancy validated our decision to apply dimensionality reduction.

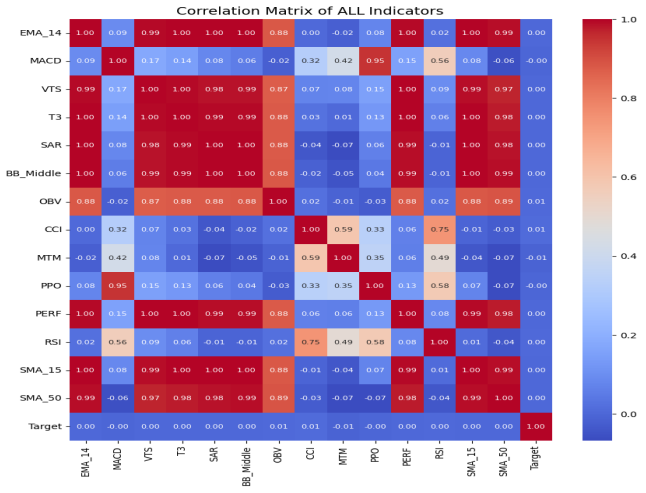

FIGURE 2. Feature Correlation Heatmap. Darker regions indicate high correlation, confirming the need for PCA to remove redundancy among the technical indicators.

D. DIMENSIONALITY REDUCTION (PCA)

To address the multicollinearity observed in the EDA phase, we applied Principal Component Analysis (PCA). We selected the number of components to retain 95% of the cumulative variance. The feature space was compressed from 19 raw dimensions to 4 Principal Components.

This transformation ensures the model trains on orthogonal (uncorrelated) signals rather than redundant noise. The 3D visualization is shown in Fig 3.

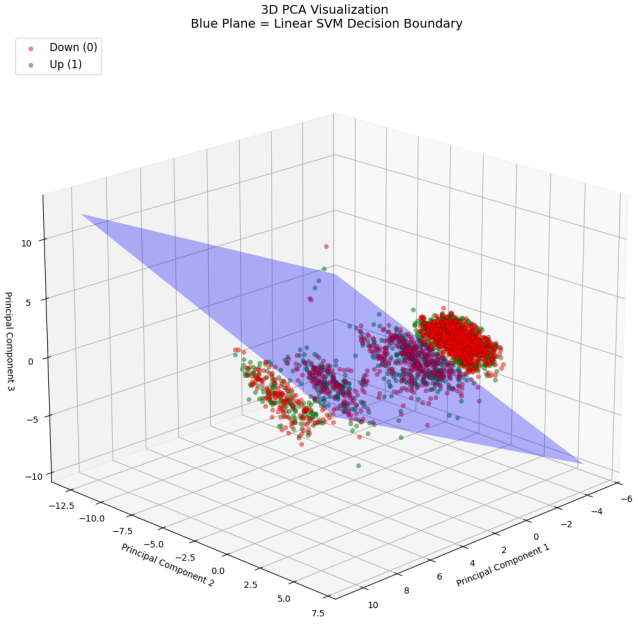

FIGURE 3. Visualizing the first 3 Principal Components. The overlap between classes (Red/Green) visually demonstrates the non-linear complexity of the prediction task.

E. PROPOSED METHOD: SOFT VOTING ENSEMBLE

We implemented five heterogeneous classifiers:

1) Logistic Regression
2) K-Nearest Neighbors
3) Support Vector Machine
4) Random Forest
5) Multi-layer Perceptron (MLP)

The proposed method integrates these into a Soft Voting Ensemble, which averages the predicted probabilities of all individual models to determine the final class, reducing variance and improving robustness.

IV. MATHEMATICAL FORMULATION OF FEATURES

To transform raw market data into predictive signals, we implemented a comprehensive set of thirteen technical indicators. Eleven were adopted from the methodology of Dey et al. [3], with the addition of Simple Moving Average (SMA) and Relative Strength Index (RSI) to capture broader market context. These features capture various market dimensions including trend, momentum, volatility, and volume. The mathematical formulations for these indicators are defined below.

Let C_t, H_t, L_t, and V_t represent the Close, High, Low, and Volume prices at time t.

A. SIMPLE MOVING AVERAGE (SMA)

The SMA calculates the arithmetic mean of prices over a specific window n. We computed this for both short-term ($n = 15$) and medium-term ($n = 50$) trends:

$$\mathrm{SMA}_t(n) = \frac{1}{n}\sum_{i=0}^{n-1} C_{t-i} \tag{2}$$

B. EXPONENTIAL MOVING AVERAGE (EMA)

Unlike SMA, the EMA places greater weight on recent data points to reduce lag. For a period $N = 14$, it is calculated as:

$$\mathrm{EMA}_t = (C_t \times \alpha) + (\mathrm{EMA}_{t-1} \times (1-\alpha)) \tag{3}$$

where the smoothing factor $\alpha = \frac{2}{N+1}$.

C. MOVING AVERAGE CONVERGENCE DIVERGENCE (MACD)

MACD is a trend-following momentum indicator derived from two EMAs.

$$\mathrm{MACDLine} = \mathrm{EMA}_{12}(C) - \mathrm{EMA}_{26}(C) \tag{4}$$
$$\mathrm{SignalLine} = \mathrm{EMA}_9(\mathrm{MACDLine}) \tag{5}$$

The divergence between these lines serves as the predictive feature.

D. TRIPLE EXPONENTIAL MOVING AVERAGE (T3)

T3 is a sophisticated smoothing technique that reduces lag better than standard EMA. We implemented the Tim Tillson algorithm using a volume factor $v = 0.7$. It is a linear combination of six EMAs (e_1 to e_6):

$$\mathrm{T3} = c_1 e_6 + c_2 e_5 + c_3 e_4 + c_4 e_3 \tag{6}$$

where coefficients are derived as:

$$\begin{aligned} c_1 &= -v^3, \quad c_2 = 3v^2 + 3v^3 \\ c_3 &= -6v^2 - 3v - 3v^3, \quad c_4 = 1 + 3v + v^3 + 3v^2 \end{aligned} \tag{7}$$

E. VOLATILITY STOP (VTS)

VTS defines the current trend using the True Range (TR). First, we calculate the Average True Range (ATR) over 14 days:

$$\mathrm{TR}_t = \max\left(H_t - L_t, |H_t - C_{t-1}|, |L_t - C_{t-1}|\right) \tag{8}$$
$$\mathrm{ATR} = \mathrm{SMA}_{14}(\mathrm{TR}) \tag{9}$$

The Volatility Stop is then derived as:

$$\mathrm{VTS} = C_t - (3 \times \mathrm{ATR}) \tag{10}$$

F. PARABOLIC SAR

The Parabolic SAR uses an iterative method to set trailing stops.

$$\mathrm{SAR}_t = \mathrm{SAR}_{t-1} + \alpha(\mathrm{EP} - \mathrm{SAR}_{t-1}) \tag{11}$$

where EP (Extreme Point) is the highest high (uptrend) or lowest low (downtrend), and α is an acceleration factor starting at 0.02 and capped at 0.2.

G. BOLLINGER BANDS (BB)

BB consists of three bands to measure volatility using a 20-day window:

$$\mathrm{MiddleBand} = \mathrm{SMA}_{20}(C) \tag{12}$$
$$\mathrm{UpperBand} = \mathrm{SMA}_{20} + (2 \times \sigma) \tag{13}$$
$$\mathrm{LowerBand} = \mathrm{SMA}_{20} - (2 \times \sigma) \tag{14}$$

where σ is the standard deviation of the close price over 20 days.

H. ON-BALANCE VOLUME (OBV)

OBV relates volume to price change. It accumulates volume on up-days and subtracts on down-days:

$$\mathrm{OBV}_t = \mathrm{OBV}_{t-1} + \begin{cases} V_t & \text{if } C_t > C_{t-1} \\ -V_t & \text{if } C_t < C_{t-1} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

I. COMMODITY CHANNEL INDEX (CCI)

CCI measures the deviation of the price from its statistical mean.

$$\mathrm{TP} = \frac{H_t + L_t + C_t}{3} \tag{16}$$
$$\mathrm{CCI} = \frac{\mathrm{TP} - \mathrm{SMA}_{14}(\mathrm{TP})}{0.015 \times \mathrm{MeanDev}} \tag{17}$$

where MeanDev is the mean absolute deviation of TP.

J. MOMENTUM (MTM)

Momentum measures the velocity of price changes. Consistent with the reference methodology, we utilized a 6-day window ($n = 6$):

$$\text{MTM} = C_t - C_{t-6} \tag{18}$$

K. PRICE OSCILLATOR (PPO)

PPO is a percentage-based version of MACD. It normalizes the difference between fast and slow EMAs:

$$\text{PPO} = \frac{\text{EMA}_{12} - \text{EMA}_{26}}{\text{EMA}_{26}} \times 100 \tag{19}$$

L. PERFORMANCE INDICATOR (PERF)

PERF measures the normalized percentage change from the beginning of the dataset:

$$\text{PERF} = \frac{C_t - C_{\text{initial}}}{C_{\text{initial}}} \times 100 \tag{20}$$

M. RELATIVE STRENGTH INDEX (RSI)

RSI identifies overbought or oversold conditions.

$$\text{RSI} = 100 - \frac{100}{1 + \text{RS}} \tag{21}$$

where

$$\text{RS} = \frac{\text{Avg Gain}_{14}}{\text{Avg Loss}_{14}} \tag{22}$$

V. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP

To prevent data leakage, we utilized Time Series Cross-Validation (TimeSeriesSplit with 3 splits). The data was split into 80% training and 20% testing sets. Standard Scaling was applied to all features before PCA to ensure that indicators with larger ranges did not dominate variance calculations.

B. PARAMETER SELECTION

Extensive Hyperparameter Tuning was performed to optimize model performance. The optimal parameters are detailed in Table 1.

TABLE 1. Parameter Selection Table

Model	Optimal Parameters
Logistic Regression	$C = 10$, penalty='l2'
K-Nearest Neighbors	$n\_neighbors = 9$, weights='uniform'
Support Vector Machine	$C = 1$, gamma='scale', kernel='rbf'
Random Forest	$max\_depth = 10$, $n\_estimators = 100$, $min\_samples\_split = 5$
MLP (Neural Net)	activation='relu', $hidden\_layer\_sizes = (50, 50)$, $alpha = 0.001$

C. MODEL PERFORMANCE METRICS

We evaluated five individual classifiers and one ensemble method. Table 2 presents the accuracy and AUC scores.

To assess the trade-off between sensitivity and specificity, we analyzed the Receiver Operating Characteristic (ROC) curves. Fig. 4 compares the curves for all five models, demonstrating that KNN (orange line) achieves the highest area under the curve.

TABLE 2. Model Performance Comparison

Model	Accuracy	AUC Score
K-Nearest Neighbors (KNN)	52.45%	0.5285
Ensemble (Voting)	52.45%	0.5256
Support Vector Machine	51.89%	0.5106
Random Forest	51.47%	0.5122
MLP Neural Network	49.93%	0.5035
Logistic Regression	48.81%	0.4916

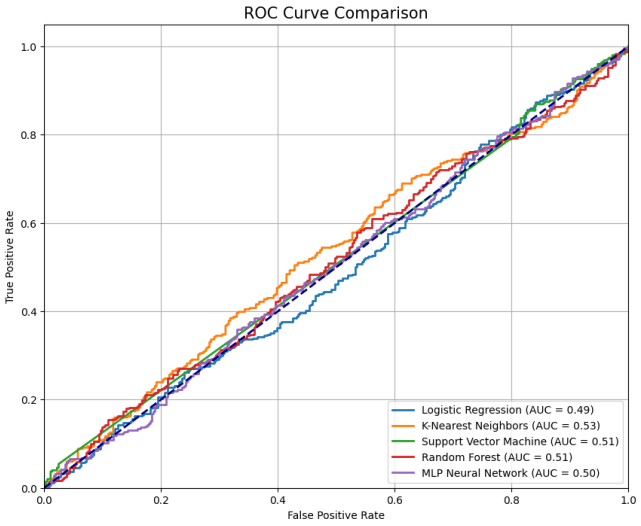

FIGURE 4. ROC Curve Comparison. KNN achieves the highest AUC of 0.5285.

To evaluate the specific types of errors made by the best-performing model, we generated a confusion matrix. Fig. 5 details the true positives and false negatives, highlighting the model's superior recall in predicting market downturns (Class 0).

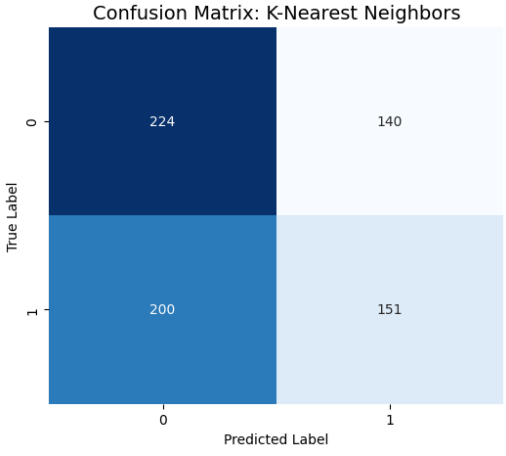

FIGURE 5. Confusion Matrix for the Best Model. The model shows a higher True Negative rate (prediction of 'Down').

D. FEATURE IMPORTANCE

To interpret the predictive power of the raw indicators, we extracted feature importance scores from a Random Forest classifier trained on the full 19-dimensional feature set prior to dimensionality reduction. To identify which technical indicators drove the model's decisions, we extracted impurity-based feature importance scores. Fig. 6 ranks the top 10 features, with Volume and Momentum indicators appearing as the most dominant signals.

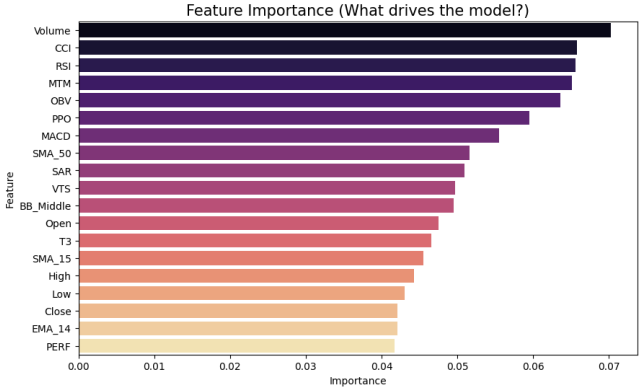

FIGURE 6. Top 10 Features impacting the model.

VI. DISCUSSION

A. PERFORMANCE ANALYSIS

The experimental results indicate that forecasting daily stock direction is a highly complex task with significant noise. K-Nearest Neighbors (KNN) and the Ensemble method outperformed the baseline, achieving an AUC of roughly 0.528. This suggests that local patterns in the feature space (exploited by KNN) are more predictive than global linear boundaries (Logistic Regression, AUC 0.49).

The Ensemble model provided stability, matching KNN's accuracy while offering a more balanced decision boundary. The confusion matrix reveals that the models are better at predicting market downturns (Recall for Class 0 $\approx$ 0.59) than upturns. This asymmetry may be due to "volatility clustering," where market drops are often sharper and more correlated than gradual rises.

B. INTERPRETATION OF TECHNICAL INDICATORS

The feature importance analysis validates several key trading theories. Trading Volume (0.070) emerged as the strongest predictor, confirming the axiom that "volume precedes price." Significant price moves are often foreshadowed by unusual volume activity.

Momentum indicators (CCI and RSI) also ranked highly, outperforming lagging trend indicators like Simple Moving Averages. This implies that for short-term daily forecasting, the *rate of change* in price is more informative than the absolute price trend itself.

C. MARKET EFFICIENCY AND LIMITATIONS

The fact that sophisticated non-linear models (SVM, MLP) hovered near 50% accuracy provides empirical support for the Efficient Market Hypothesis (EMH) in its semi-strong form. The marginal edge gained (approx. 2.5% over random chance) suggests that while some inefficiency exists, it is difficult to exploit using technical data alone. The primary limitation of this study is the exclusion of fundamental data (earnings, news), which often drives the major trend shifts that technical indicators fail to anticipate.

VII. CONCLUSION

This project presented a machine learning framework for forecasting stock market trends using 13 technical indicators. By employing a rigorous methodology involving PCA for dimensionality reduction and a Voting Ensemble for classification, we achieved a predictive accuracy of 52.45%. While the accuracy is modest, the study validates two critical financial concepts:

1) Volume Precedes Price: Feature importance analysis confirmed that Trading Volume is the strongest predictor of future movement.
2) Market Noise: The difficulty in significantly beating the 50% baseline confirms the high-noise nature of daily stock data.

DATA AND CODE AVAILABILITY

The source code used for all experiments and constructed for this study, are publicly available at: https://github.com/FarazHeydar/Forecasting-Stock-Market-Trends.

REFERENCES

[1] Z. Wang, Z. Hu, F. Li, S. B. Ho and E. Cambria, "Learning-based stock trending prediction by incorporating technical indicators and social media sentiment," *Cognitive Computation*, vol. 15, no. 3, pp. 1092-1102, 2023.
[2] K. Fu and Y. Zhang, "Incorporating Multi-Source Market Sentiment and Price Data for Stock Price Prediction," *Mathematics*, vol. 12, no. 10, p. 1572, 2024.
[3] P. P. Dey, N. Nahar and B. M. M. Hossain, "Forecasting Stock Market Trend using Machine Learning Algorithms with Technical Indicators," *International Journal of Information Technology and Computer Science*, vol. 12, no. 3, pp. 32-38, 2020.

FARAZ HEYDAR was born in Tehran, Iran, in 1998. He received the B.S. degree in Computer Engineering - Software from the Islamic Azad University (Central Tehran Branch), Tehran, Tehran, Iran, in 2022. He is currently pursuing the M.Sc. degree in computer science at Lakehead University, Thunder Bay, ON, Canada.

His research interests include artificial intelligence, software engineering, machine and deep learning, and computer vision.

HESAMEDDIN BITARAFAN RAJABI was born in Tehran, Iran, in 1998. He received the B.S. degree in Computer Engineering (Software) from the University of Science and Culture, Tehran, in 2022. He is currently pursuing the M.Sc. degree in Computer Science at Lakehead University, Thunder Bay, ON, Canada.

His academic interests include artificial intelligence, computer vision, machine learning, and related areas of intelligent systems.

SAAD BIN AHMED received the master's degree in intelligent systems from Technische Universitaet, Kaiserslautern, Germany, in 2012, and the Ph.D. degree in computer sciences from Universiti Teknologi Malaysia, in 2019.,He has served as a Lecturer with King Saud bin Abdulaziz University for Health Sciences (KSAU-HS), Riyadh, Saudi Arabia, for seven years. He served as a Research Assistant with the Image Understanding and Pattern Recognition Research Group, Technische Universitaet. He is currently an Assistant Professor with Lakehead University, Thunder Bay, ON, Canada. He has accomplished his postdoctoral research from the University of Western Ontario, London, Canada. He is also the Director of the Image Analysis and Pattern Identification Research Laboratory (IAPI-RL). He is associated with various computer science journals and provides his expertise in reviewing novel contributions. By utilizing his extensive academic experience, he has significantly contributed which has been published as research articles in impact factor journals, conferences, and book chapters. His research interests include document image analysis, machine learning, computer vision, and optical character recognition. He is in the image analysis and pattern recognition field for 15 years and has been involved in various pioneer researches like collection of handwritten Urdu data and used it for Urdu character recognition. He provided his expertise in capturing Arabic scene text images and performing research on collected samples by machine learning and pattern recognition techniques. The focus of his research is on hyperspectral image analysis and explainable AI.