# A Hybrid Relational Approach Toward Stock Price Prediction and Profitability

Manali Patel, Krupa Jariwala, and Chiranjoy Chattopadhyay

***Abstract*—An accurate estimation of future stock prices can help investors maximize their profits. The current advancements in the area of artificial intelligence (AI) have proven prevalent in the financial sector. Besides, stock market prediction is difficult owing to the considerable volatility and unpredictability induced by numerous factors. Recent approaches have considered fundamental, technical, or macroeconomic variables to find hidden complex patterns in financial data. At the macro level, there exists a spillover effect between stock pairs that can explain the variance present in the data and boost the prediction performance. To address this interconnectedness defined by intrasector stocks, we propose a hybrid relational approach to predict the future price of stocks in the American, Indian, and Korean economies. We collected market data of large-, mid-, and small-capitalization peer companies in the same industry as the target firm, considering them as relational features. To ensure efficient feature selection, we have utilized a data-driven approach, i.e., random forest feature permutation (RF2P), to remove noise and instability. A hybrid prediction module consisting of temporal convolution and linear model (TCLM) is proposed that considers irregularities and linear trend components of the financial data. We found that RF2P-TCLM gave the superior performance. To demonstrate the real-world applicability of our approach in terms of profitability, we created a trading method based on the predicted results. This technique generates a higher profit than the existing approaches.**

***Impact Statement*—The dependency of stock prices on numerous factors makes prediction difficult. The previous approaches focused on stock-specific indicators, neglecting the interconnection between stock pairs. This study addresses this factor by incorporating relational dimensions, as well as temporality, into the prediction task. A hybrid prediction module is proposed to capture irregularities and trend patterns efficiently. The robustness of this approach is proven by profitability analysis.**

***Index Terms*—Feature optimization, hybrid model, relational modeling, stock prediction, trading strategy.**

## I. Introduction

STOCK trading has seen massive growth recently due to higher short-term returns, inflation protection, and safe long-term benefits compared with other investment options. With accurate forecasting, one can maximize their profit by buying stocks that are anticipated to rise in the future and selling those that are likely to fall. But at the same time, stock prices are highly volatile and unpredictable in nature, making them a complex dynamic system to model [1]. The dependency of stock prices on numerous factors such as political, global, and economic situations and the psychology of investors is the reason behind its volatility [2]. According to the efficient market hypothesis (EMH) [3], prices follow a random walk pattern that reflects all available information, making forecasting difficult. Although the recent advancement of artificial intelligence (AI) in predicting the stock market has proven efficient and has given rise to a new interdisciplinary domain called financial technology (FinTech) [4], [5]. It has attracted researchers from the business and computer science communities to develop an efficient forecasting model.

There have been two mainstream approaches to analyze market: fundamental and technical. Fundamentalists conduct an in-depth quantitative and qualitative study of a stock, examining balance sheets, profit and loss statements, and accounting data to determine its intrinsic value. A buy or sell recommendation is given by comparing the stock's intrinsic value with the market price [6], [7]. Technical analysis looks at time series data to uncover intricate hidden patterns. To extract profitable patterns, many lagging and leading technical indicators are developed [8], [9]. These indicators are fed into the prediction model as raw features, as they contain relevant information about price patterns. The use of technical indicators as features introduces subjective bias and necessitates optimization to improve prediction efficiency.

To forecast future prices, the current methods solely consider fundamental and technical indicators along with market data. However these features are inadequate to elucidate the interdependencies among stock pairs. In real life, there exists the spillover effect between the stock pairs that are connected through some relationship [10], [11]. The movement in one's stock price also affects the other associated stocks and can provide insight into the overall market situation. To account for this, we propose a relational approach that assesses market dynamics by studying the spillover effect between the same sector stock pairs. The intrasector dynamics provide a valuable source of information, as stocks within the same sector exhibit a higher correlation with each other and tend to move in a similar manner.

Manuscript received 21 May 2024; accepted 26 May 2024. Date of publication 31 May 2024; date of current version 12 November 2024. *(Corresponding author: Chiranjoy Chattopadhyay.)*

Manali Patel and Krupa Jariwala are with the Department of Computer Science and Engineering, Sardar Vallabhbhai National Institute of Technology, Surat 395007, India (e-mail: d21co004@coed.svnit.ac.in; knj@coed.svnit.ac.in).

Chiranjoy Chattopadhyay is with the School of Computing and Data Sciences, FLAME University, Pune 412115, India (e-mail: chiranjoy.chattopadhyay@flame.edu.in).

This article has supplementary downloadable material available at https://doi.org/10.1109/TAI.2024.3408129, provided by the authors.

Digital Object Identifier 10.1109/TAI.2024.3408129

According to the capitalization values, stocks are classified into three categories, i.e., large, mid, and small. The mid and small-cap stocks are considered to be more volatile and sensitive to external events compared with the large-cap stocks [12]. This volatility can be a useful indicator for studying the price fluctuation of a target large-cap stock. To capture this, we have considered the closing price of peer companies (large, mid, and small) within the same sector as relational features. We have also included market and technical indicators to build a comprehensive set of features. A hybrid model, RF2P-TCLM is proposed to optimize the features and capture nonlinear and linear trend patterns to predict future prices. The key contributions of the proposed work are

1) A relational approach that captures extensive relationships among intrasector stock pairs.
2) Creation of a new dataset comprising large-, mid-, and small-cap peer companies from India, the USA, and Korea.
3) A novel RF2P inspired approach to find the optimal set of features.
4) An amalgamation of temporal convolution network (TCN) and a linear model (LM) is proposed to account for the linear and nonlinear components in stock prices.
5) Real-world validation of the proposed approach by developing a trading strategy and yielding higher profitability.

The article is organized as follows: Section II reviews the various approaches to stock market forecasting. Section III formulates the research question and explains the proposed methodology. Section IV gives an overview of the dataset and evaluation parameters. Section V represents the results and discussion. Section VI presents the applicability of the model, followed by the ablation study in Section VII. The conclusion is presented in Section VIII.

## II. Related Work

To forecast future prices, the prediction methods are mainly classified into three categories: statistical, deep learning (DL), and hybrid models.

### A. Statistical Models

The statistical models apply time series modeling to forecast the future stock market. Mashadihasanli [13] applied the ARIMA (3,1,5) model to predict the future price of the Istanbul and Turkiye stock markets. Idress et al. [14] also used ARIMA to forecast the future stock price of the National Stock Exchange (NSE) index of India. Promma and Chutsagulprom [15] proposed a vector autoregressive model, AVAR-KF with the Kalman filter for parameter estimation. Another statistical approach, generalized auto regressive conditional heteroskedasticity (GARCH), is used to predict the market's future crash events [16] and volatility [17].

The statistical models work under the assumption of data stationarity, which does not hold for financial data. Moreover, these models fail to forecast long-term trends due to their inherent model complexities and parametric nature.

### B. DL Models

Over the years, research works [18], [19] reveal that stock prices move in a nonlinear fashion and that DL models have an edge over statistical models. These models can extract features on their own and learn complex nonlinear patterns, giving them a competitive advantage in stock market predictions [20].

State-of-the-art vision models have been employed to predict the stock market due to their ability to capture local patterns efficiently using kernel techniques. Hoseinzade and Haratizadeh [21] proposed a unique approach, a 2-D and 3-D CNNPred model to examine the correlation among various market indices to predict future prices. Macroeconomic variables such as the return rate of different market indices and commodity prices, along with technical indicators, were considered. The analysis suggested that interconnections among market indices have an effect on prediction accuracy. Patel et al. [22] considered the impact of peer companies in the same business to forecast the future price movement direction of stocks listed on the National Stock Exchange of India. A correlation analysis was performed to identify large-cap companies related to the target firm.

To capture both long-term and short-term trends in the CSI 300 index, Yang and Jing [23] proposed the Bidirectional long short-term memory (Bi-LSTM) model. It outperformed the vanilla LSTM, SVR, and ARIMA models. Wu et al. [24] proposed a generative adversarial networks (GANs) with piecewise linear representation approach to predict trading actions. It outperformed the LSTM model. Lee et al. [25] fused stock prices, macroeconomic variables, and month and day of the week information to predict future price movements. The multimodal features were extracted using the multihead self-attention mechanism. Sáenz et al. [26] applied a clustering algorithm based on quarterly financial reports to find similar stocks. Based on the clustering information, deep LSTM and ARIMA models were trained.

### C. Hybrid Models

There have been attempts to create hybrid models that take advantage of various single models. Touzani and Douzi [27] proposed a hybrid technique that combined LSTM and GRU to develop a trading strategy for the Moroccan market. Shah et al. [28] presented a hybrid CNN-LSTM model to forecast the NIFTY-50 index's future closing price. They proposed a time-distributed layer to capture the temporal dependency. Lin et al. [29] proposed an ensemble machine learning framework to predict the stock trend using candlestick chart patterns. Xiao and Su [30] used a combination of statistical and deep approaches, ARIMA-LSTM to forecast the closing price of the S & P 500 index. The linear component of this model is calculated using the ARIMA model, and the nonlinear part is simulated using the LSTM model, with errors from both models included to further update the weights of these models.

Wu et al. [31] proposed a SACLSTM model that considered the effect of leading indicators such as future and option data on the prediction performance of selected stocks from the USA and Taiwan. The features were extracted using the CNN model and fed into the LSTM model. Zhang et al. [32] combined three

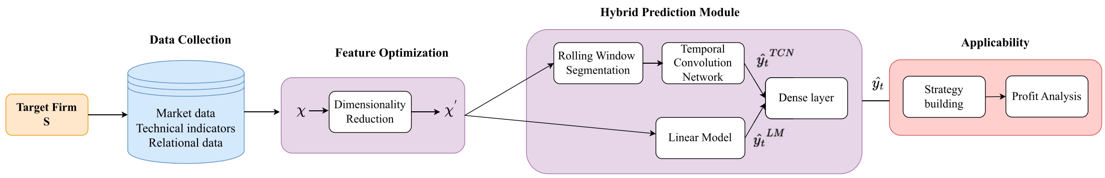


Fig. 1. Schematic representation of the proposed methodology.

models: a convolution neural network, bidirectional LSTM, and an attention mechanism to predict the price of the Chinese stock index CSI 300. Shaban et al. [33] proposed an SMP-DL model that combined LSTM and bidirectional GRU models to predict the future price of the different stocks.

There have been approaches to optimize the parameters of prediction models using meta-heuristic algorithms that search for a global solution to minimize the error. Zheng et al. [34] proposed the BA-SVR model to optimize the free parameters of the SVM model in the short, medium, and long term to forecast several Chinese market indices. Farahani and Hajiagha [35] presented an integrated meta-heuristic-based ANN model to forecast multiple stock indices. To select the appropriate technical indicators, a genetic algorithm (GA) is used. Corizzo and Rozen [36] proposed an ensemble approach that combined stock prices and news headlines to predict future price movements and make informed trading decisions.

From the literature, we found that hybrid models have an advantage over single models. Moreover, the financial data exhibits linear and nonlinear behavior. To capture this, we propose a hybrid prediction model, RF2P-TCLM to capture both aspects. The recent work has concentrated on considering technical indicators as additional input characteristics along with market data to build the prediction models. To demonstrate the importance of interconnectedness and effectively learn market dynamics, we introduced relational features that establish linkages between other stocks in the same sector.

TABLE I
FEATURE DESCRIPTION

| Feature | Name | Description |
|---|---|---|
| Market ($F_{\text{Market}}$) | Open price ($f_o$) | The price at which the first trade occurred. |
| | Close price ($f_c$) | The last traded price of a share. |
| | Low price ($f_l$) | The lowest price of a trade. |
| | High price ($f_h$) | The highest price of a trade. |
| | Volume ($f_v$) | The number of shares traded. |
| Technical ($F_{\text{Technical}}$) | Return ($ti_{\text{return}}$) | $\frac{\text{close}_{t-1} - \text{close}_t}{\text{close}_{t-1}}$ |
| | Simple moving average ($ti_{\text{SMA20}}$) | $\frac{\text{close}_1 + \text{close}_2 \ldots. + \text{close}_{20}}{20}$ |
| | Exponential moving average ($ti_{\text{EMA20}}$) | $[\text{close}_t - \text{EMA}_{t-1}] * \alpha + \text{EMA}_{t-1}$ |
| | Average directional index ($ti_{\text{ADX}}$) | $+DI = \frac{\Sigma_{t=1}^{14} DM - \frac{\Sigma_{t=1}^{14} DM14}{+} \text{Current DM}}{\text{Average True Range}} * 100$<br>$-DI = \frac{\Sigma_{t=1}^{14} DM - \frac{\Sigma_{t=1}^{14} DM}{14} - \text{Current DM}}{\text{Average True Range}} * 100$<br>$DX = \frac{\lvert +DI - -DI \rvert}{\lvert +DI + -DI \rvert} * 100$<br>$\text{ADX} = \frac{(\text{Prior ADX} * 13) + \text{Current ADX}}{14}$<br>+DM = current high - previous high<br>-DM = previous low - current low |
| | Moving average convergence divergence ($ti_{\text{MACD}}$) | 12 period EMA $-$ 26 period EMA |
| | Bollinger bands ($ti_{BB}$) | $B\_MA =$ Moving Average$(TP, n)$<br>$BU = B\_MA +$ no. of standard deviation $* \sigma$<br>$BL = B\_MA -$ no. of standard deviation $* \sigma$<br>$TP = (\text{High} + \text{Low} + \text{Close})/3$ |
| | Relative Strength Index ($ti_{RSI}$) | $100 - \left( \frac{100}{1 + \frac{\text{14 day average gain}}{\text{14 day average loss}}} \right)$ |
| | Stochastic %K ($ti_{S\%k}$) | $\%K = \frac{\text{close} - \text{low}(t-14)}{\text{high}(t-14) - \text{low}(t-14)} * 100$ |
| | Stochastic %D ($ti_{S\%D}$) | $\%D = 3 -$ period moving average of $\%K$ |
| | Williams %R ($ti_{W\%R}$) | $\frac{\text{high}(t-14) - \text{close}}{\text{high}(t-14) - \text{low}(t-14)}$ |
| | Commodity Channel Index $ti_{\text{CCI}}$ | $\frac{TP - n \text{ period average of } TP}{0.015 * \text{Mean Deviation}}$ |
| Relational ($F_{\text{Relational}}$) | Closing price of Large, Mid, and Small cap peer companies $\{r_1, r_2, \ldots., r_r\}$ | https://www.spglobal.com<br>https://www.nasdaq.com<br>https://www.nseindia.com<br>https://www.moneycontrol.com<br>https://www.tickertape.in<br>https://global.krx.co.kr |

## III. Methodology

In this section, we first formulate the research problem mathematically and explain our proposed framework in detail.

1) **Problem Formulation:** Given a series of historical prices of stock $S$ consisting of $N$ data points denoted as: $X_S = \{x_1, \ldots\ldots, x_{N-1}, x_N\}$, where $x_i \in \mathbb{R}^d$, our goal is to predict the future price $\hat{y_{t+1}}$. $d$ represents the dimension of input features. This single-step price prediction problem is formulated as follows:

$$\hat{y_{t+1}} = G(X_S^{(t)}). \tag{1}$$

$G(.)$ defines the proposed prediction framework, which is demonstrated in Fig. 1. It has four main components: 1) Data collection: to construct a comprehensive set of features; 2) Feature optimization: which selects the rich and meaningful features; 3) Prediction module: a hybrid approach consisting of a nonlinear DL model and an LM followed by dense layers to combine the prediction results;4) Applicability: that demonstrates the applicability of this approach, a strategy is constructed using the predicted value, and the profit earned is determined. Next, all modules are discussed in detail.

### A. Data Collection

We have considered the stocks from different sectors of three economies, i.e., America, India, and Korea. The market, technical, and relational data of each stock is considered to build a comprehensive set of features. These features are illustrated in Table I.

To identify large, mid, and mall capitalization peer companies belonging to the same industry as the target firm, we utilized official websites of considered stock exchanges and stock-related platforms, as mentioned in Table I. These extracted peer companies' closing prices are considered a relational feature. The cardinality of each feature set is in the order of $|F_{\text{relational}}| > |F_{\text{Technical}}| > |F_{\text{market}}|$.

Our objective is to study the effect of market, technical, and relational features on the target firm's closing price to predict

**Algorithm 1** Random Forest Feature Permutation (RF2P) feature selection.

**Input:** $X_S = \{x_1, x_2, ........, x_{N-1}, x_N\} \in \mathbb{R}^{N\times d}$,
$Y_S = \{f_{c_1}, f_{c_2}, ....., f_{c_N}\}, \Theta$ {N is total data points, $f_c$ is closing price, and $\Theta$: User defined threshold.}
**Output:** $X^{'}_S = \{x^{'}_1, x^{'}_2, ........, x^{'}_{N-1}, x^{'}_N\} \in \mathbb{R}^{N\times d'}$
$(D_{train} : X_{train}, Y_{train}), (D_{valid} : X_{valid}, Y_{valid}) \leftarrow split(X_S, Y_S)$
$m \leftarrow RandomForest(X_{train}, Y_{train})$
$s \leftarrow Score(m, (X_{valid}, Y_{valid}))$
$F^{'} = \phi$
**for** each feature j in F **do**
  **for** each repetition k in 1,2,....,K **do**
    Randomly shuffle j to generate corrupted version $\tilde{D}_{valid(k,j)}$
    $s_{k,j} \leftarrow Score(m, \tilde{D}_{valid(k,j)})$
  **end for**
  $importance_j \leftarrow s - \frac{1}{K}\Sigma_{k=1}^{K} s_{k,j}$
  **if** $importance_j \geq \Theta$ **then**
    $F^{'} = F^{'} \cup j$
  **end if**
**end for**
**return** $F^{'}$

the future price. We construct an initial input feature set denoted as $F = \{F_{\text{Market}} - \{f_c\} \cup F_{\text{Technical}} \cup F_{\text{Relational}}\}$ by excluding the closing price feature and perform feature optimization to reduce the noise and instability. Here, $|F| = d$.

### B. Feature Optimization

Feature optimization aims to select the most important features and eliminate redundant ones to improve the efficiency of the model. We have adapted RF2P as a feature optimization approach. A feature matrix of stock $S$ consisting of $N$ data points denoted as $X_S \in \mathbb{R}^{N\times d}$ and a predefined threshold value $\Theta$ is given as an input. This approach is described as:

*1) Random Forest Feature Permutation (RF2P):* A feature permutation is a model-agnostic approach that studies the effect of each feature on the considered model's (here, random forest) performance and generates an importance score. RF is an efficient ensemble machine learning method that considers the decisions of multiple trees to reach a single result [37]. This RF-based feature selection approach has been effective for various tasks [38], [39]. To produce new predictions, feature permutation takes a fitted RF model and randomly shuffles predictors. A feature importance score is determined by comparing new predictions to predictions generated by a random forest classifier on the data. If the prediction performance improves after shuffling, these features are retained; otherwise, they are removed. The procedure is explained in Algorithm 1.

The RF2P generates the importance score for each feature present in the feature set $F$. It becomes essential to select meaningful features that can explain the variability present in the data. A reduced number of features might degrade performance, whereas too many features can increase the complexity of a model. To balance this out, we have to choose the threshold value in such a manner that it can cover a significant number of features. The feature importance range for RF2P changes depending on the data distribution, and we excluded features with negative feature permutation values. We have selected two threshold values for RF2P: >10e-07 and >10e-06. If a feature importance score calculated by RF2P is greater than the threshold value, that feature is added to the optimal feature set $F^{'}$, where $F^{'} \subset F$. The optimized feature matrix of stock $S$ denoted as $X^{'}_S = \{x^{'}_1, x^{'}_2, ....., x^{'}_N\} \in \mathbb{R}^{N\times d^{'}}$, where $d^{'} = |F^{'}|$ and $d^{'} < d$ is given to the prediction module. To make notations clear, we remove the subscript $S$.

### C. Prediction Module

The aim of this module is to identify hidden linear and nonlinear patterns that can explain the present price fluctuations. The financial data has mainly three components: Irregularity, Trend, and Temporality. To capture irregularities, a deep model is adapted that extracts nonlinear patterns exhibited by multiple features. An LM is utilized to capture long-term movements. The temporality defines dependency on the previous closing price values, and it is handled differently by both models. The optimized feature matrix $X^{'} \in \mathbb{R}^{N\times d^{'}}$ is given to the nonlinear and LM, which generates the prediction for the next day.

*1) Nonlinear Model:* This module identifies the complex nonlinear interactions between various features and predicts the future. Sequential models such as recurrent neural networks (RNNs), long short term memory (LSTM), and gated recurrent unit (GRU) are commonly used to predict future stock prices. However, they suffer from exploding and vanishing gradients and are incapable of capturing long-term reliance. A variation of convolution neural network known as temporal convolution network (TCN) [40] is adapted to capture the nonlinearity. It is designed to capture the local and temporal patterns in the sequential data with its large receptive fields. It has outperformed state-of-the-art sequential models in a variety of tasks, including vision [41], language processing [42], [43], and multiple time series forecasting tasks [44], [45].

To ensure temporality, we construct the feature set of a nonlinear model by concatenating the closing price feature of a target firm: $F_{\text{NLM}} = \{F^{'} \cup f_c\}$, where $|F_{\text{NLM}}| = d' + 1$. A rolling window segmentation is performed on the updated feature matrix $X^{\text{NLM}} = \{x_1^{\text{NLM}}, x_2^{\text{NLM}}, ....., x_N^{\text{NLM}}\} \in \mathbb{R}^{N\times(d^{'}+1)}$ to transform a data point at a time $t$ that can encode previous $P$ data points. In the beginning, $P$ data points depicting the lookback window are selected as training points, and according to the forecast horizon, $H$ data points are considered test data. In the next iterations, we roll forward for $H$ data points and continue splitting the dataset where $P$ data points are considered to predict the next $H$ data points. In our problem settings, we have considered the forecast horizon $H$ as 1. The visual representation is shown in Fig. 2.

The time-dependent series at instance $t$ consists of $P$ previous data points and is obtained using rolling window segmentation. It is denoted as $\chi_t^{\text{TCN}} = \{x_{t-P+1}^{\text{NLM}}, ......x_{t-1}^{\text{NLM}}, x_t^{\text{NLM}}\} \in \mathbb{R}^{P\times(d'+1)}$ is given to the TCN model.

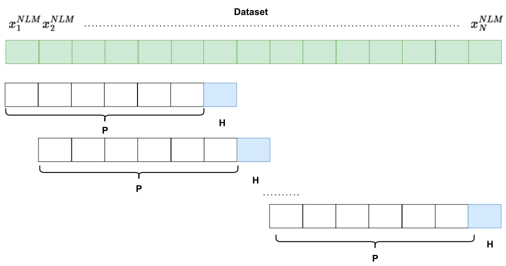


Fig. 2. Illustration of the rolling window segmentation technique.

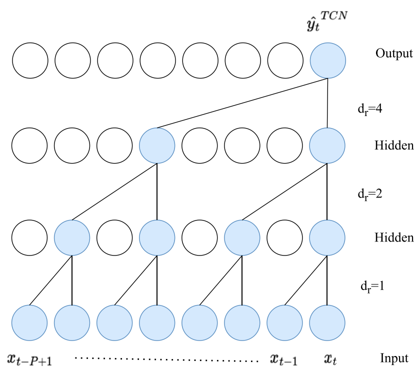


Fig. 3. Illustration of the TCN Architecture having two hidden layers with different dilation rates denoted with $d = 1, 2, 4$ and a kernel size of 2.

The architecture of TCN is made up of causal 1-D convolution and dilated layers with the same input and output lengths. The causal layer ensures the exclusive dependence of present input on previous time steps and prevents leakage of future information. To acquire information over a lengthy period of time, we require a large deep model, or kernel size, which results in a complex model. To avoid this, a 1-D dilated layer is utilized to represent long-range dependency by expanding the receptive field. At each hidden layer, the output $\hat{y_t}^{\text{TCN}}$ is calculated as follows:

$$\hat{y_t}^{\text{TCN}} = \sum_{i=0}^{k-1} f(i) \cdot x_{t-(d_r \cdot i)}^{\text{NLM}} \tag{2}$$

where filter is denoted as $f : \{0, 1, 2, ...k-1\} \rightarrow \mathbb{R}$, $k$ is filter size, $d_r$ is dilation rate, and $x_t^{\text{NLM}}$ is a feature vector at time instance $t$. The architecture of TCN is shown in Fig. 3.

We have trained a nonlinear model with dilation rates $\{1, 2, 4, 16, 32\}$ with 64 filters of size 2.

*2) LM:* The aim of this module is to capture the temporal dependency and the effect of the multiple features on the target firm's price patterns. To ensure temporal dependency, we define a feature vector of the closing price of the target firm for previous $P$ data points denoted as $\{f_{c_t}, f_{c_{t-1}}, ...., f_{c_{t-P+1}}\}$. This along with the optimized feature vector at time instance $t$ denoted as $x_t^{'} \in \mathbb{R}^{d^{'}}$ obtained from the previous layer, is given to the below model to predict the future price $\hat{y_t}^{LM}$ as follows:

$$y_t^{\hat{LM}} = [\alpha_1, \alpha_2, ...\alpha_{d'}] \begin{bmatrix} x_{f_1} \\ x_{f_2} \\ \vdots \\ x_{f_{d'}} \end{bmatrix} + [\beta_1, \beta_2, ....., \beta_P] \begin{bmatrix} f_{c_t} \\ f_{c_{t-1}} \\ \vdots \\ f_{c_{t-P+1}} \end{bmatrix} + \gamma_0. \tag{3}$$

TABLE II
DESCRIPTION OF THE DATASETS USED FOR EXPERIMENTATION

| Economy | Industry | Company | Ticker | No. of Peer Companies |
|---|---|---|---|---|
| India | Information Technology | Tata Consultancy Services | TCS | 25 |
| | Healthcare | Sun Pharmaceutical | SUNPHARMA | 46 |
| | Financial | HDFC Bank Ltd | HDFCBANK | 25 |
| | Fast moving consumer goods | Hindustan Unilever Ltd | HINDUNILVR | 7 |
| | Oil, gas and consumable fuels | Reliance Industries Ltd | RELIANCE | 11 |
| America | Computer software | Alphabet Inc Class C | GOOG | 50 |
| | Medical/dental instruments | Abbott Laboratories | ABT | 36 |
| | Major banks | Citigroup Inc | C | 84 |
| | Package goods/cosmetics | Procter & Gamble Co | PG | 29 |
| | Integrated oil companies | Exxon Mobil Corp | XOM | 25 |
| Korea | Software development and supply | NCsoft Corporation | 036570.KS | 10 |
| | Manufacture of petroleum products | S-oil | 010950.KS | 11 |
| | Manufacture of food products | Samyang food | 003230.KS | 16 |
| | Manufacture of medicaments | YUHAN | 000100.KS | 34 |
| | Insurance | HANWHA LIFE | 088350.KS | 79 |

The $\alpha$, $\beta$, and $\gamma$ parameters are learnt during the training process. The prediction results obtained by the linear and nonlinear models are given to the nonlinear approximation function $F(.)$ to make the final prediction $\hat{y_t}$ as follows:

$$\hat{y_t} = F(y_t^{\hat{LM}}, y_t^{\hat{\text{TCN}}}). \tag{4}$$

The $F(.)$ consists of a dense layer with 32 neurons, followed by the final prediction layer. The prediction module is trained with the Adam optimizer for 100 epochs with a batch size of 16. We denote the prediction module combined with Random Forest Feature Permutation as RF2P-TCLM.

## IV. Experimentation Details

In this section, we discuss the experimental setup, which comprises a dataset, evaluation metrics, and baseline models.

### A. Dataset

To prove the generalizability of the proposed approach, we selected five stocks from different sectors of three economies—India, America, and Korea for examination, as shown in Table II. We have collected historical data of considered companies and peer companies from 2012 to 2023 for experimentation. The number of large, mid, and small capitalization peer companies having trading activity during 2012–2023 is also depicted in Table II which constructs the set of relational features. The list of peer companies is provided in the supplementary material.

80% is used for training, and the rest 20% is used for testing. The training set contains 2356, 2404, and 2352 data points, and the test set contains 589, 600, and 587 data points for Indian, American, and Korean economies, respectively.

### B. Baseline Models

We have considered the baseline methods presented in Table III to compare our proposed work. The same feature set and model architecture described in their respective works are implemented.

### C. Evaluation Parameters

As our objective is to predict future stock prices based on historical data, we have used metrics designed for time-dependent

TABLE III
DETAILS OF BASELINE MODELS

| Model | Features | Feature Optimization | Prediction Module |
|---|---|---|---|
| [35] | Market data, technical indicators | Genetic Algorithm | GA-ANN |
| [32] | Market data, return | - | A hybrid CNN-BiLSTM model with attention mechanism |
| [23] | Market data, return | - | BiLSTM |
| [28] | Market data, technical indicators, exchange rates | - | A hybrid CNN-LSTM and time distributed layers |
| [22] | Relational data of large cap peer companies | Correlation | LSTM |
| Ours | Market data, technical data, relational data | PCA with 90% and 99% threshold value | The proposed TCLM module |

data. We have considered the following evaluation parameters to emphasize correctness of the prediction and directional accuracy.

1) **Mean Absolute Error (MAE):** It measures the average absolute difference between the actual and predicted values and is calculated as follows:
$$\text{MAE} = \frac{\Sigma_{t=1}^{n}|y_t - \hat{y}_t|}{n}. \tag{5}$$
2) **Root Mean Squared Error (RMSE):** RMSE calculates the square root of the squared difference between the real and predicted values. It penalizes larger errors. It is calculated as follows:
$$\text{RMSE} = \sqrt{\frac{\Sigma_{t=1}^{n}|y_t - \hat{y}_t|^2}{n}}. \tag{6}$$
3) **Mean Absolute Percentage Error (MAPE):** MAPE is a relative error measure that calculates the average percentage difference between actual and predicted values. It is calculated as follows:
$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}|\frac{y_t - \hat{y}_t}{y_t}|. \tag{7}$$
4) **Mean Directional Accuracy (MDA):** MDA focuses on the direction of price movements. It measures how often the predicted direction (upward or downward) aligns with the actual direction of movement. It ranges from 0% to 100%. It is calculated as follows:
$$\text{MDA} = \frac{\Sigma_{t=2}^{n}\text{sign}(y_t - y_{t-1}) * \text{sign}(\hat{y}_t - \hat{y}_{t-1})}{n-1} \tag{8}$$
$$\text{sign}(y_t - y_{t-1}) = \begin{cases} 1, & \text{if } y_t > y_{t-1} \\ -1, & \text{otherwise} \end{cases} \tag{9}$$

where $n$, $y_t$, and $\hat{y}_t$ represent the number of test data points, actual, and predicted values respectively. The lower value of error matrices i.e., MAE, RMSE, MAPE is preferred. The lower values indicate better accuracy in predictions. On the other hand, a higher value of MDA is preferred, as it reflects the model's ability to accurately predict future movements.

## V. RESULT AND DISCUSSION

In this section, we present the prediction results obtained by the proposed RF2P-TCLM and baseline models. As mentioned in Section III-B1, we have considered two different threshold values with RF2P (>10e-06 and >10e-07) to build the optimal feature set. The threshold value >10e-07 will cover a larger number of features compared with >10e-06. As the prediction module is the same, this will give us insight into the degree of feature optimization on the prediction performance. We experimented with a lag size ($P$) of 15 to investigate the significance of the temporal dimension defined in Section III. The results obtained are discussed in the next section.

### A. Indian Stock Market

In this section, we present the prediction results for the companies considered for the Indian stock market. Table IV represents the results of the Indian stocks for the lag value 15. It is evident that the proposed RF2P-TCLM has given superior performance compared to other baseline models, as the value of error metrics (MAE, RMSE, and MAPE) is lower and the MDA value is higher for all considered Indian stocks.

As illustrated in Table IV, threshold values (>10e-07 and >10e-06) have a significant effect on the prediction performance. Apart from the MAE, RMSE, and MAPE scores, RF2P-TCLM has also shown superior performance in predicting the directional movements for all the Indian stocks. In terms of MDA, the RF2P-TCLM has shown 15%, 5%, 13%, 7%, and 10% improvement for TCS, SUNPHARMA, RELIANCE, HINDUNILVR, and HDFCBANK, respectively.

For the Indian stocks, these threshold values have achieved superior performance: (TCS: RF2P-TCLM (>10e-07)), (SUNPHARMA: RF2P-TCLM (>10e-06)), (RELIANCE: RF2P-TCLM (>10e-06)), (HINDUNILVR: RF2P-TCLM (>10e-07)), (HDFCBANK: RF2P-TCLM (>10e-07)).

### B. American Stock Market

In this section, we present the prediction results for the companies considered for the American stock market. Table V represents the results of the American stocks for the lag values of 15. It is evident that, for stocks GOOG, PG, XOM, and C, the RF2P-TCLM has outperformed other models as the value of error matrices (MAE, RMSE, and MAPE) is lower and the MDA value is higher for these stocks. For ABT stock, PCA-TCLM (90%) has given superior performance compared with the RF2P-TCLM model with a 0.54 MAE, 0.71 RMSE, and 0.49 MAPE value. As the prediction module is the same for both models, we can confirm that the features selected by the RF2P are not adequate compared with PCA to extract the temporal patterns of ABT stock. However, for MDA value, RF2P-TCLM (>10e-07) has given the highest performance with 85% directional accuracy for ABT stock.

The RF2P-TCLM has achieved a lower error value (GOOG: 1.02), (PG: 0.28), (XOM: 0.41), and (C: 0.92) for stocks in terms of MAPE compared with other models. This signifies that the prediction results obtained by the RF2P-TCLM model are closer to the actual values. It has also obtained a higher MDA value for all considered American stocks.

For the American stock market, the combination of these feature optimization approaches and threshold values has achieved the highest performance: (GOOG: RF2P-TCLM (>10e-07)),

TABLE IV
ANALYSIS OF INDIAN STOCKS FOR LAG VALUE OF 15

| **Company** | **TCS** | | | | **SUNPHARMA** | | | | **RELIANCE** | | | | **HINDUNILVR** | | | | **HDFCBANK** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** |
| GA-ANN [35] | 50.89 | 64.50 | 1.48 | 48.46 | 22.26 | 25.48 | 2.38 | 48.29 | 84.21 | 94.15 | 3.41 | 53.06 | 47.04 | 57.14 | 1.89 | 51.53 | 26.38 | 33.34 | 1.73 | 50.85 |
| CNN-BiLSTM-AM [32] | 87.29 | 114.10 | 2.44 | 52.55 | 23.32 | 32.19 | 2.34 | 46.59 | 112.38 | 130.01 | 4.43 | 53.41 | 60.24 | 74.75 | 2.35 | 51.36 | 26.60 | 34.51 | 1.71 | 52.38 |
| Bi-LSTM [23] | 77.86 | 89.32 | 2.24 | 51.36 | 15.99 | 20.38 | 1.62 | 47.49 | 71.58 | 83.81 | 2.85 | 52.72 | 74.63 | 82.66 | 2.97 | 50.68 | 21.80 | 27.95 | 1.44 | 51.19 |
| CNN-LSTM [28] | 137.13 | 157.30 | 3.97 | 49.65 | 17.71 | 25.41 | 1.82 | 47.61 | 80.64 | 94.55 | 3.25 | 51.02 | 55.00 | 69.27 | 2.24 | 50.51 | 61.24 | 70.25 | 3.92 | 52.72 |
| SM2PNet [22] | 41.76 | 53.53 | 1.22 | 47.11 | 14.41 | 18.58 | 1.48 | 51.19 | 98.72 | 108.79 | 3.95 | 51.19 | 37.03 | 45.77 | 1.51 | 54.76 | 28.32 | 34.86 | 1.83 | 51.21 |
| PCA-TCLM (90%) | 28.01 | 33.83 | 0.81 | 71.42 | 3.68 | 4.68 | 0.39 | 82.31 | 36.29 | 41.20 | 1.46 | 75.34 | 17.64 | 22.28 | 0.71 | 73.97 | 9.67 | 12.33 | 0.63 | 76.71 |
| PCA-TCLM (99%) | 142.61 | 157.48 | 4.14 | 64.62 | 3.38 | 4.56 | 0.35 | 81.80 | 32.65 | 39.69 | 1.31 | 74.14 | 14.19 | 18.69 | 0.57 | 77.72 | 25.31 | 41.69 | 1.61 | 70.91 |
| RF2P-TCLM (>10e-07) | **12.81** | **16.54** | **0.37** | **82.48** | 8.57 | 14.01 | 0.82 | 85.54 | 9.74 | 12.91 | 0.39 | **85.03** | **8.02** | **10.53** | **0.32** | 82.99 | **5.12** | 6.74 | **0.33** | **84.18** |
| RF2P-TCLM (>10e-06) | 30.09 | 40.13 | 0.88 | 80.27 | **2.87** | **3.62** | **0.31** | **86.56** | **9.61** | **12.27** | **0.38** | 84.35 | 8.27 | 10.86 | 0.33 | **83.33** | 5.21 | **6.72** | 0.34 | 82.48 |

Note: Bold values indicate the best performance.

TABLE V
ANALYSIS OF AMERICAN STOCKS FOR LAG VALUE OF 15

| **Company** | **GOOG** | | | | **PG** | | | | **XOM** | | | | **ABT** | | | | **C** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** |
| GA-ANN [35] | 6.41 | 8.32 | 5.05 | 47.00 | 2.23 | 2.83 | 1.52 | 49.66 | 3.22 | 3.85 | 3.37 | 50.16 | 2.00 | 2.56 | 1.81 | 45.66 | 0.81 | 1.04 | 1.55 | 50.83 |
| CNN-BiLSTM-AM [32] | 4.13 | 5.20 | 3.23 | 50.5 | 5.78 | 6.47 | 3.83 | 50.51 | 2.37 | 3.28 | 2.44 | 53.00 | 2.51 | 3.41 | 2.14 | 48.51 | 0.77 | 1.02 | 1.47 | 54.33 |
| Bi-LSTM [23] | 4.98 | 5.94 | 4.23 | 48.83 | 2.07 | 2.49 | 1.39 | 49.83 | 1.58 | 2.02 | 1.73 | 52.83 | 1.79 | 2.24 | 1.64 | 47.51 | 0.71 | 0.95 | 1.38 | 52.66 |
| CNN-LSTM [28] | 7.23 | 8.65 | 6.14 | 47.51 | 3.31 | 4.49 | 2.27 | 50.00 | 6.21 | 6.89 | 6.51 | 50.16 | 6.05 | 8.48 | 5.73 | 49.16 | 2.07 | 2.43 | 4.31 | 54.66 |
| SM2PNet [22] | 2.22 | 2.89 | 1.85 | 51.83 | 3.81 | 1.95 | 1.02 | 48.16 | 1.91 | 2.42 | 2.00 | 52.00 | 2.85 | 3.34 | 2.64 | 48.83 | 0.80 | 1.03 | 1.56 | 53.33 |
| PCA-TCLM (90%) | 2.27 | 3.01 | 1.83 | 65.51 | 0.92 | 1.12 | 0.61 | 84.83 | 0.78 | 0.98 | 0.85 | 83.33 | **0.54** | **0.71** | **0.49** | 81.83 | 1.97 | 2.24 | 3.77 | 57.16 |
| PCA-TCLM (99%) | 1.36 | 1.73 | 1.11 | 73.83 | 0.69 | 0.84 | 0.46 | 83.66 | 1.52 | 2.07 | 1.46 | 86.33 | 0.66 | 0.87 | 0.58 | 81.51 | 1.23 | 1.44 | 2.63 | 90.51 |
| RF2P-TCLM (>10e-07) | **1.28** | **1.67** | **1.02** | **89.66** | 1.45 | 1.85 | 0.96 | 82.66 | **0.39** | **0.54** | **0.41** | **89.00** | 0.74 | 0.92 | 0.65 | **85.00** | 0.61 | 0.66 | 1.21 | 92.33 |
| RF2P-TCLM (>10e-06) | 1.65 | 1.84 | 1.41 | 89.51 | **0.41** | **0.53** | **0.28** | **88.51** | 1.01 | 1.65 | 1.05 | 85.16 | 2.79 | 3.49 | 2.65 | 77.66 | **0.46** | **0.50** | **0.92** | **92.51** |

Note: Bold values indicate the best performance.

TABLE VI
ANALYSIS OF KOREAN STOCKS FOR LAG VALUE OF 15

| **Company** | **036570.KS** | | | | **010950.KS** | | | | **000100.KS** | | | | **088350.KS** | | | | **003230.KS** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** | **MAE** | **RMSE** | **MAPE** | **MDA** |
| GA-ANN [35] | 29.13 | 37.11 | 6.38 | 48.81 | 5.93 | 6.38 | 6.72 | 50.17 | 1.25 | 1.88 | 2.21 | 49.65 | 7.60 | 0.99 | 2.85 | 49.82 | 7.33 | 9.56 | 5.89 | 48.46 |
| CNN-BiLSTM-AM [32] | 11.19 | 16.97 | 2.55 | 48.21 | 1.91 | 2.40 | 2.18 | 51.71 | 1.24 | 2.02 | 2.06 | 46.92 | 6.17 | 0.79 | 2.23 | 50.34 | 4.85 | 9.53 | 3.17 | 47.61 |
| Bi-LSTM [23] | 10.38 | 16.35 | 2.33 | 51.02 | 1.46 | 1.97 | 1.63 | 47.26 | 0.73 | 1.18 | 1.29 | 47.00 | 5.45 | 0.72 | 1.98 | 48.29 | 3.41 | 5.12 | 2.73 | 46.07 |
| CNN-LSTM [28] | 34.40 | 40.92 | 8.31 | 49.31 | 1.97 | 2.59 | 2.27 | 47.78 | 2.45 | 3.05 | 4.51 | 46.92 | 14.12 | 1.87 | 5.38 | 49.65 | 5.37 | 9.03 | 3.85 | 46.75 |
| SM2PNet [22] | 10.56 | 16.89 | 2.39 | 47.09 | 2.36 | 3.28 | 2.73 | 49.31 | 1.03 | 1.43 | 1.85 | 47.78 | 5.81 | 0.79 | 2.14 | 49.14 | 4.34 | 5.88 | 3.51 | 47.09 |
| PCA-TCLM (90%) | **3.22** | **6.43** | **0.69** | 87.03 | 0.49 | 0.65 | 0.55 | 89.41 | 0.42 | 0.73 | 0.72 | 79.86 | 4.87 | 0.55 | 1.82 | 86.00 | 1.63 | 2.62 | 1.24 | 78.15 |
| PCA-TCLM (99%) | 4.98 | 7.21 | 1.14 | **88.91** | **0.37** | **0.49** | **0.41** | 90.78 | 0.41 | 0.66 | 0.71 | 81.39 | 6.01 | 0.67 | 2.22 | 89.07 | 3.70 | 7.98 | 2.24 | 76.45 |
| RF2P-TCLM (>10e-7) | 4.31 | 6.60 | 1.21 | 86.34 | 1.36 | 1.70 | 1.52 | 89.59 | **0.23** | **0.41** | **0.39** | **82.76** | 4.37 | 0.53 | 1.66 | 89.41 | **1.24** | 2.34 | **0.84** | **83.95** |
| RF2P-TCLM (>10e-6) | 6.39 | 7.89 | 1.78 | 87.37 | 0.58 | 0.66 | 0.69 | **91.29** | 0.33 | 0.61 | 0.55 | 82.42 | **2.18** | **0.27** | **0.82** | **90.11** | 1.29 | **2.03** | 0.97 | 82.76 |

Note: Bold values indicate the best performance.

(PG: RF2P-TCLM(>10e-06)), (XOM: RF2P-TCLM (>10e-07)), (ABT: PCA-TCLM (90%)), and (C: RF2P-TCLM (>10e-06)).

## C. Korean Stock Market

In this section, we present the prediction results for the companies considered for the Korean stock market. Table VI represents the results of the Korean stocks for the lag value of 15[1].

For stocks 000100.KS, 088350.KS, and 003230.KS, the RF2P-TCLM has outperformed other models for all parameters. For 036570.KS and 010950.KS stocks, PCA-TCLM (90%) and PCA-TCLM (99%) have given superior performance compared with the RF2P-TCLM model for MAE, RMSE, and MAPE parameters. The features selected by the PCA are more meaningful compared with the RF2P approach for these stocks. However, RF2P-TCLM (>10e-06) has outperformed the PCA-TCLM (99%) performance for the MDA parameter with a 91.29% value for 010950.KS stock.

In terms of matrices (MAE, RMSE, and MAPE), the RF2P-TCLM has achieved lower prediction errors except for the stocks 036570.KS and 010950.KS. In terms of MDA, the RF2P-TCLM has shown 5%, 2%, 1%, and 7% improvement for 010950.KS, 000100.KS, 088350.KS, and 003230.KS, respectively. For 036570.KS stock, PCA-TCLM has shown an improvement of 2% over the RF2P-TCLM approach.

For the Korean stocks, the combination of these feature optimization approaches and threshold values has achieved the highest performance: (036570.KS: PCA-TCLM (90%)), (010950.KS: PCA-TCLM (99%)), (000100.KS: RF2P-TCLM (>10e-07)), (088350.KS: RF2P-TCLM (>10e-06)), (003230.KS: RF2P-TCLM (>10e-07)).

Based on the above analysis, we confirm that the hybrid model RF2P-TCLM, which considers relational features, has outperformed the other approaches for a large number of stocks. We can observe the significant difference between the performance of RF2P-TCLM and baseline models for MAE, RMSE, MAPE, and MDA evaluation parameters. Furthermore, the MDA parameter's superior performance indicates that the model's directional predictions match the actual trends. This demonstrates the robustness of the proposed RF2P-TCLM model in predicting future stock prices and directional movements. Moreover, the ability of the RF2P-TCLM model to perform efficiently across different stocks with distinct data distributions proves its generalizability.

One possible reason behind the failure of baseline models GA-ANN [35] and Bi-LSTM [23] can be the consideration of the insufficient features for the prediction task as well as using a single model. Even the hybrid models with rich feature sets

[1]The magnitude of the MAE and RMSE is normalized for Korean stocks by a factor of $10^3$ for easier comparison.

TABLE VII
PROFIT ANALYSIS BY THE MACD-RSI-BB STRATEGY FOR INDIAN, AMERICAN, AND KOREAN STOCKS

| Stock ID | GA-ANN [35] | CNN-BiLSTM-AM [32] | Bi-LSTM [23] | CNN-LSTM [28] | SM2PNet [22] | Buy-and-Hold | PCA-TCLM | RF2P-TCLM |
|---|---|---|---|---|---|---|---|---|
| TCS | 3092.00 | 1814.00 | 2568.00 | 2323.79 | 2807.00 | 460.00 | 2919.73 | **3194.00** |
| SUNPHARMA | 951.00 | 613.00 | 895.00 | **1171.29** | 711.00 | 553.00 | 967.02 | 880.00 |
| RELIANCE | 2056.00 | 1247.00 | 1842.00 | 2051.00 | 2253.00 | 404.00 | 2523.70 | **2577.00** |
| HINDUNILVR | 1920.00 | 1325.00 | 1522.00 | 1511.00 | 1437.00 | 161.00 | 1883.08 | **2081.00** |
| HDFCBANK | 1316.00 | 822.00 | 1231.00 | 1061.00 | 1406.00 | 211.00 | 1452.18 | **1501.00** |
| GOOG | 26.51 | 13.99 | 14.68 | 18.12 | 19.85 | -3.12 | 17.85 | **26.99** |
| PG | 3.88 | 2.74 | 6.28 | 0.00 | 6.78 | 6.03 | 8.85 | **7.68** |
| XOM | 26.50 | 25.12 | 31.52 | 33.06 | 28.66 | 42.58 | 34.01 | **33.71** |
| ABT | 113.43 | 79.16 | 61.16 | **127.22** | 84.99 | -14.58 | 107.17 | 101.34 |
| C | 50.93 | 41.20 | 41.31 | 38.04 | 48.91 | -18.56 | **65.59** | 58.84 |
| 036570.KS | 588 237.29 | **499 597.80** | 375 729.16 | 470 457.79 | 347 498.27 | -568 000.00 | 428 279.55 | 424 793.32 |
| 010950.KS | 120 073.37 | 80 705.90 | 125 615.30 | 123 804.84 | 84 614.90 | -27 500.00 | 129 006.37 | **129 434.24** |
| 000100.KS | 67 248.25 | 54 421.14 | 72 364.10 | 73 931.86 | 60 679.73 | 9591.62 | 67 762.07 | **79 158.45** |
| 003230.KS | 258 717.06 | 221 799.52 | 228 762.19 | 215 818.90 | 240 259.73 | 118 800.00 | 263 821.88 | **264 056.99** |
| 088350.KS | 4671.98 | 4632.75 | 4092.79 | 3612.17 | 3585.68 | -715.00 | 4553.91 | **4875.30** |

Note: Bold values indicate the best performance.

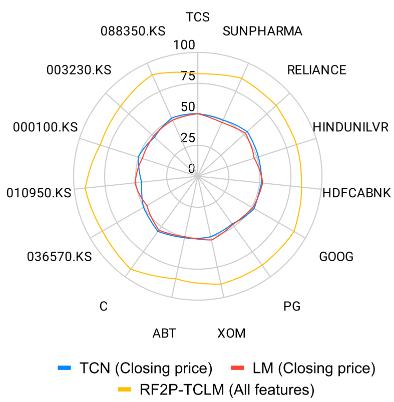


Fig. 4. Comparison between TCN, LM, and RF2P-TCLM models.

consisting of market and diverse technical indicators (CNN-BiLSTM-AM [32] and CNN-LSTM [28]) did not improve the prediction results. One reason for this could be the neglect of the associations between stock pairs, which can more effectively explain external events than these stock-specific parameters. This aspect is addressed in this work by considering the relational features of large, mid, and small-capitalization peer companies. Another reason can be negligence in capturing both the nonlinear and linear components, which add valuable information about the trend and irregularities, respectively.

Moreover, proper feature selection plays an important role in the stock movement prediction task. The SM2PNet [22] achieved poor performance as a static parameter correlation coefficient was used to optimize the features. The PCA-TCLM gave the second best results after the RF2P-TCLM. The drawback of the PCA feature optimization approach is its sensitivity to outliers, resulting in the incorrect projection of the data. We confirm that the ensemble machine learning-based approach RF2P has proven efficient. A hybrid prediction module is capable of capturing the underlying trend, irregularity, and temporality of financial data.

## VI. PROFITABILITY ANALYSIS

In this section, we demonstrate the model's relevance in real-world applications by developing a trading strategy and calculating the profit. Traders utilize various combinations of technical indicators for short-term trading to increase earnings. We provide the prediction results obtained by the baseline models and the proposed RF2P-TCLM model, denoted as Data($D$) as an input to apply a strategy and calculate the profit. For each company, the threshold value that has achieved the highest directional accuracy is considered, as mentioned in Section V. To produce buy or sell recommendations, we investigated employing a combination of trend (moving average convergence divergence), momentum (relative strength index), and volatility (Bollinger bands) technical indicators. The formula used to calculate these indicators is mentioned in Table I. Algorithm 2 explains how to construct a buy or sell signal that integrates all three technical indicators. A buy or sell signal is generated when the combined signal strength is greater than the threshold $\delta$ (30%).

**Algorithm 2** MACD-RSI-BB Strategy Implementation.

**Input:** Data $D$, $W = [w_{macd} : 1, w_{rsi} : 3, w_{bb} : 2]$, Signal strength $\delta$ {$W$ represents weight given to each indicator.}
**Output:** Buy/Sell signal for $D$.

```
signal ← [] {An array to store buy/sell signal.}
probability ← [] {An array to store signal strength.}
for each datapoint i in D do
    score ← 0
    for each w in W do
        score ← score + D_i * w
    end for
    score ← score / sum(W)
    if score ≤ δ then
        signal[i] ← Buy
        probability[i] ← score * (1)
    else
        signal[i] ← Sell
        probability[i] ← score * (−1)
    end if
end for
return signal
```

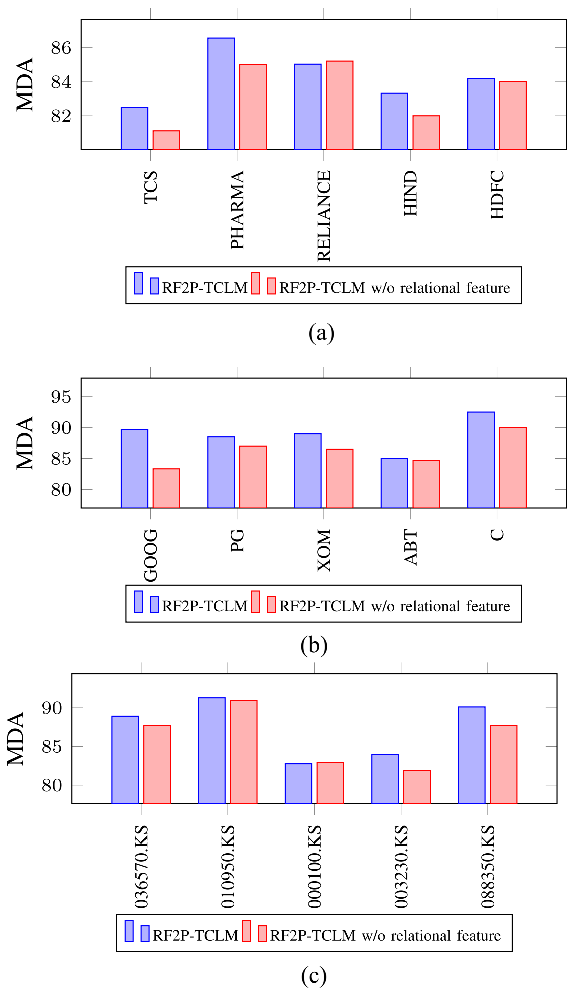


Fig. 5. Comparison between RF2P-TCLM and RF2P-TCLM w/o relational features for the (a) Indian; (b) American; (c) Korean economies.

From the generated buy and sell signal, trading is done. It follows the principle of the “buy low, sell high theory”. Following this, we first buy stock at a lower price and then sell later at a higher price. A complete transaction is considered when we encounter both “BUY” and “SELL” signals. The trading follows these rules

1) If consecutive signals are “BUY”, then we buy that particular stock at a minimum price and sell when we encounter a next sell signal.
2) If consecutive signals are “SELL”, then we sell that particular stock at a maximum price to book a profit and buy when we encounter a next buy signal.

We have considered an initial budget of Rs. 5000, 200 Dollar, and 240 000 Won for Indian, American, and Korean stocks, respectively. We have considered a trade cost of Rs. 20 (Dollar or Won) for each transaction. We calculate the profit by taking the difference between the accumulated amount from all trades and the initial investment amount during the test period. We calculate the profit generated by the above-mentioned MACD-RSI-BB strategy using the prediction results obtained by the baseline and proposed RF2P-TCLM models. Moreover, we compare our trading strategy with the buy-and-hold approach. This technique purchases stock at a specific point in time and holds it for an extended period of time. Table VII represents the profit generated by each prediction model and buy-and-hold approach.

It is evident that the RF2P-TCLM has achieved the maximum profit for the majority of stocks except for SUNPHARMA, ABT, C, and 036570.KS. We also observe that by applying a thoughtful strategy to the obtained prediction, we can enhance our profits instead of relying on a buy-and-hold approach. During the test phase, American stocks (GOOG, ABT, and C) and Korean stocks (036570.KS, 010950.KS, and 088350.KS) have a bearish market condition, resulting in a loss if we implement a buy and hold strategy. Instead of that, using the MACD-RSI-BB strategy, it is able to generate a profit. This represents the robustness of our prediction model and trading strategy.

TABLE VIII
ANALYSIS ON EXECUTION TIME

| Model | Execution Time (s) |
|---|---|
| GA-ANN [35] | 1100 |
| CNN-BiLSTM-AM [32] | 400 |
| Bi-LSTM [23] | 1000 |
| CNN-LSTM [28] | 900 |
| SM2PNet [22] | 500 |
| PCA-TCLM | 300 |
| RF2P-TCLM | 300 |

## VII. ABLATION STUDY

In this section, we analyze the model from two perspectives, i.e., module and execution time analysis.

### A. Module Analysis

To highlight the importance of the proposed prediction model (RF2P-TCLM), which considers a comprehensive set of features and a hybrid approach to make predictions, we trained the individual nonlinear (TCN) and LMs presented in Section III for all the considered stocks. The TCN and LM models are trained with the closing price feature. We considered the threshold values for which RF2P-TCN has achieved the best performance reported in Section V. We evaluate these models' performance based on the value of the MDA parameter.

The outer, middle, and inner layers in Fig. 4 show the MDA values of TCN, LM, and RF2P-TCLM models, respectively.

We can observe the significant difference between the directional accuracy achieved by the TCN, LM, and RF2P-TCLM models. The nonlinear TCN and LM are able to achieve directional accuracy in the range of [45–55%] when trained with the closing price feature. This implies that the market data is not sufficient to explain the volatility of the financial data. The technical and relational features contain valuable information about the price trend patterns, helping to achieve better accuracy. Moreover, these models alone are not sufficient to capture the complex nature of the stock market. The ability of these models is enhanced by considering the comprehensive set of features and the hybrid model that efficiently fuses the nonlinear and linear components.

Furthermore, to prove the effectiveness of relational features, we trained the proposed RF2P-TCLM prediction model without relational features for all the considered stocks. The result on the MDA parameter is shown in Fig. 5.

From Fig. 5, we observe that the RF2P-TCLM has outperformed the RF2P-TCLM trained without relational features for the majority of stocks. For RELIANCE and 000100.KS stocks, the RF2P-TCLM trained without relational features has given better performance compared with the RF2P-TCLM. However, for these stocks, the difference between the performance of both models is negligible. This justifies the importance of the proposed relational features.

### B. Execution Time Analysis

In this section, we evaluate the performance of the prediction models based on the execution time. Table VIII presents the analysis of execution time (in seconds) taken by each model considered in this work. The execution time taken by the RF2P-TCLM and PCA-TCLM is the same, as they share the same prediction model. Compared with other approaches, RF2P-TCLM has taken a lesser amount of time.

## VIII. Conclusion

The task of forecasting stock prices presents a significant challenge for researchers due to its inherent intricacy. This study presents a novel and comprehensible methodology for forecasting stock prices in the different stocks of the American, Indian, and Korean economies. Our approach, RF2P-TCLM combines relational dependency analysis with hybrid prediction techniques. The methodology employed in our study integrates financial market data, technical indicators, and relational features that were previously inaccessible. Our model outperforms baseline models in predicting future prices and directional movements using the hybrid TCLM and data-driven feature optimization technique. The inclusion of relational features substantially improves the performance, and the implementation of a refined trading strategy results in greater financial gains when compared with a passive buy-and-hold approach.

## References


[1] P. Yang and X. Hou, "Research on dynamic characteristics of stock market based on big data analysis," *Discrete Dyn. Nature Soc.*, vol. 2022, no. 1, pp. 1–8, Mar. 2022.

[2] S. Li, Y. Wang, Z. Zhang, and Y. Zhu, "Research on the factors affecting stock price volatility," in *Proc. 7th Int. Conf. Financial Innov. Econ. Develop.*, China: Atlantis Press, 2022, pp. 2884–2889.

[3] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *J. Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[4] F. G. D. C. Ferreira, A. H. Gandomi, and R. T. N. Cardoso, "Artificial intelligence applied to stock market trading: A review," *IEEE Access*, vol. 9, pp. 30898–30917, 2021.

[5] S. Mokhtari, K. K. Yen, and J. Liu, "Effectiveness of artificial intelligence in stock market prediction based on machine learning," *Int. J. Comput. Appl.*, vol. 183, no. 7, pp. 1–8, Jun. 2021.

[6] M. X. Hanauer, M. Kononova, and M. S. Rapp, "Boosting agnostic fundamental analysis: Using machine learning to identify mispricing in European stock markets," *Finance Res. Lett.*, vol. 48, Aug. 2022, Art. no. 102856.

[7] F. Wang, X. S. Yan, and L. Zheng, "Do sophisticated investors follow fundamental analysis strategies? Evidence from hedge funds and mutual funds," *Rev. Accounting Stud.*, vol. 29, no. 2, pp. 1097–1146, Mar. 2023.

[8] L. Almeida and E. Vieira, "Technical analysis, fundamental analysis, and Ichimoku dynamics: A bibliometric analysis," *Risks*, vol. 11, no. 8, Aug. 2023, Art. no. 142.

[9] S. Dongrey, "Study of market indicators used for technical analysis," *Int. J. Eng. Manage. Res.*, vol. 12, no. 2, pp. 64–83, 2022.

[10] U. Ali and D. Hirshleifer, "Shared analyst coverage: Unifying momentum spillover effects," *J. Financial Econ.*, vol. 136, pp. 649–675, Oct. 2019.

[11] J. Li, L. Cheng, X. Zheng, and F.-Y. Wang, "Analyzing the stock volatility spillovers in Chinese financial and economic sectors," *IEEE Trans. Comput. Social Syst.*, vol. 10, no. 1, pp. 269–284, Feb. 2023.

[12] S. Jena, A. Tiwari, A. Dash, and E. J. A. Abakah, "Volatility spillover dynamics between large-, mid-, and small-cap stocks in the time-frequency domain: implications for portfolio management," *J. Risk Financial Manage.*, vol. 14, no. 11, Nov. 2021, Art. no. 531.

[13] T. Mashadihasanli, "Stock market price forecasting using the Arima model: an application to Istanbul, Turkiye," *J. Econ. Policy Res.*, vol. 9, no. 2, pp. 439–454, 2022.

[14] S. M. Idrees, M. A. Alam, and P. Agarwal, "A prediction approach for stock market volatility based on time series data," *IEEE Access*, vol. 7, pp. 17287–17298, 2019.

[15] N. Promma and N. Chutsagulprom, "Forecasting financial and macroeconomic variables using an adaptive parameter VAR-KF Model," *Math. Comput. Appl.*, vol. 28, no. 1, Feb. 2023, Art. no. 19.

[16] E. Koo and G. Kim, "A hybrid prediction model integrating GARCH models with a distribution manipulation strategy based on LSTM networks for stock market volatility," *IEEE Access*, vol. 10, pp. 34743–34754, 2022.

[17] R. Mattera and P. Otto, "Network log-arch models for forecasting stock market volatility," *Int. J. Forecasting,* Jan. 2024, doi: 10.1016/j.ijforecast.2024.01.002.

[18] K. Olorunnimbe and H. Viktor, "Deep learning in the stock market—A systematic survey of practice, backtesting, and applications," *Artif. Intell. Rev.*, vol. 56, no. 3, pp. 2057–2109, Jun. 2023.

[19] M. M. Kumbure, C. Lohrmann, P. Luukka, and J. Porras, "Machine learning techniques and data for stock market forecasting: A literature review," *Expert Syst. Appl.*, vol. 197, 2022, Art. no. 116659.

[20] A. W. Li and G. S. Bastos, "Stock market forecasting using deep learning and technical analysis: A systematic review," *IEEE Access*, vol. 8, pp. 185232–185242, 2020.

[21] E. Hoseinzade and S. Haratizadeh, "CNNpred: CNN-based stock market prediction using a diverse set of variables," *Expert Syst. Appl.*, vol. 129, pp. 273–285, Sep. 2019.

[22] M. Patel, K. Jariwala, and C. Chattopadhyay, "SM2PNet: A deep learning approach towards Indian stock market movement prediction," in *Proc. 8th Int. Conf. Convergence Technol. (I2CT)*, 2023, pp. 1–7.

[23] Y. Mo and J. Wang, "Adaptability of financial time series prediction based on BiLSTM," *Procedia Comput. Sci.*, vol. 199, pp. 18–25, 2022.

[24] J.-L. Wu, X.-R. Tang, and C.-H. Hsu, "A prediction model of stock market trading actions using generative adversarial network and piecewise linear representation approaches," *Soft Comput.*, vol. 27, no. 12, pp. 8209–8222, 2022.

[25] T.-W. Lee, P. Teisseyre, and J. Lee, "Effective exploitation of macroeconomic indicators for stock direction classification using the multimodal fusion transformer," *IEEE Access*, vol. 11, pp. 10275–10287, 2023.

[26] J. V. Sáenz, F. M. Quiroga, and A. F. Bariviera, "Data vs. information: Using clustering techniques to enhance stock returns forecasting," *Int. Rev. Financial Anal.*, vol. 88, Jul. 2023, Art. no. 102657.

[27] Y. Touzani and K. Douzi, "An LSTM and GRU based trading strategy adapted to the Moroccan market," *J. Big Data*, vol. 8, no. 1, pp. 1–16, 2021.

[28] A. Shah, M. Gor, M. Sagar, and M. Shah, "A stock market trading framework based on deep learning architectures," *Multimedia Tools Appl.*, vol. 81, no. 10, pp. 1–19, Apr. 2022.

[29] Y. Lin, S. Liu, H. Yang, and H. Wu, "Stock trend prediction using candlestick charting and ensemble machine learning techniques with a novelty feature engineering scheme," *IEEE Access*, vol. 9, pp. 101433–101446, 2021.

[30] D. Xiao and J. Su, "Research on stock price time series prediction based on deep learning and autoregressive integrated moving average," *Sci. Program.*, vol. 2022, no. 1, pp. 1–12, 2022.

[31] J. M.-T. Wu, Z. Li, N. Herencsar, B. Vo, and C.-W. Lin, "A graph-based CNN-LSTM stock price prediction algorithm with leading indicators," *Multimedia Syst.*, vol. 29, pp. 1751–1770, Feb. 2021.

[32] J. Zhang, L. Ye, and Y. Lai, "Stock price prediction using CNN-BiLSTM-attention model," *Mathematics*, vol. 11, no. 9, 2023, Art. no. 1985.

[33] W. M. Shaban, E. Ashraf, and A. E. Slama, "SMP-DL: A novel stock market prediction approach based on deep learning for effective trend forecasting," *Neural Comput. Appl.*, vol. 36, no. 4, pp. 1849–1873, 2024.

[34] J. Zheng, Y. Wang, S. Li, and H. Chen, "The stock index prediction based on SVR model with bat optimization algorithm," *Algorithms*, vol. 14, no. 10, 2021, Art. no. 299.

[35] M. S. Farahani and S. H. R. Hajiagha, "Forecasting stock price using integrated artificial neural network and metaheuristic algorithms compared to time series models," *Soft Comput.*, vol. 25, no. 13, pp. 8483–8513, 2021.

[36] R. Corizzo and J. Rosen, "Stock market prediction with time series data and news headlines: A stacking ensemble approach," *J. Intell. Inf. Syst.*, vol. 62, no. 1, pp. 27–56, 2023.

[37] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[38] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, 2020, Art. no. 52.

[39] K. O. Nti, A. Adekoya, and B. Weyori, "Random forest based feature selection of macroeconomic variables for stock market prediction," *Am. J. Appl. Sci.*, vol. 16, no. 7, pp. 200–212, 2019.

[40] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Proc. Comput. Vis. (ECCV) Workshops*, The Netherlands: Springer, 2016, pp. 47–54.

[41] Y. Zhao et al., "A temporal-aware relation and attention network for temporal action localization," IEEE Trans. Image Process., vol. 31, pp. 4746–4760, 2022.

[42] J. Ren, W. Wu, G. Liu, Z. Chen, and R. Wang, "Bidirectional gated temporal convolution with attention for text classification," *Neurocomputing*, vol. 455, pp. 265–273, Sep. 2021.

[43] S. Rathor and S. Agrawal, "Sense understanding of text conversation using temporal convolution neural network," *Multimedia Tools Appl.*, vol. 81, no. 7, pp. 9897–9914, Mar. 2022.

[44] Y. Lin, I. Koprinska, and M. Rana, "Temporal convolutional attention neural networks for time series forecasting," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Piscataway, NJ, USA: IEEE Press, 2021, pp. 1–8.

[45] W. Tian and W. Li, "Multi-mode spatial-temporal convolution network for traffic flow forecasting," in *Proc. 2nd Int. Conf. Big Data Inf. Educ. (ICBDIE)*, 2021, pp. 278–281.