Journal of King Saud University - Computer and Information Sciences 36 (2024) 102162

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Full Length Article

Integrating PCA with deep learning models for stock market Forecasting: An analysis of Turkish stocks markets

Taner Uçkan

Van Yüzüncü Yıl University, Department of Computer Engineering, Van 65000, Turkey

ARTICLE INFO

Keywords:
BIST
Deep learning
Feature Selection
Technical indicators
PCA
Stock price prediction

ABSTRACT

Financial data such as stock prices are rich time series data that contain valuable information for investors and financial professionals. Analysis of such data is critical to understanding market behaviour and predicting future price movements. However, stock price predictions are complex and difficult due to the intense noise, non-linear structures, and high volatility contained in this data. While this situation increases the difficulty of making accurate predictions, it also creates an important area for investors and analysts to identify opportunities in the market. One of the effective methods used in predicting stock prices is technical analysis. Multiple indicators are used to predict stock prices with technical analysis. These indicators formulate past stock price movements in different ways and produce signals such as buy, sell, and hold. In this study, the most frequently used ten different indicators were analyzed with PCA (Principal Component Analysis. This study aims to investigate the integration of PCA and deep learning models into the Turkish stock market using indicator values and to assess the effect of this integration on market prediction performance. The most effective indicators used as input for market prediction were selected with the PCA method, and then 4 different models were created using different deep learning architectures (LSTM, CNN, BiLSTM, GRU). The performance values of the proposed models were evaluated with MSE, MAE, MAPE and R2 measurement metrics. The results obtained show that using the indicators selected by PCA together with deep learning models improves market prediction performance. In particular, it was observed that one of the proposed models, the PCA-LSTM-CNN model, produced very successful results.

1. Introduction

The stock market is an investment area where people can buy and sell shares of companies that go public. Predicting stock prices is of great importance for investors to evaluate opportunities and for financial companies to reduce risks arising from credit-based transactions. Stock markets have a very complex structure depending on many factors. Due to this complexity, collecting information from various sources and using it efficiently in order to more effectively predict stock price fluctuations provides a great advantage in terms of the sector(Ma et al., 2023). Most investors looking to invest in the stock market are primarily concerned about the direction in which a stock price might move. Therefore, they seek sources that can provide preliminary information about the movements of stock prices. In predicting stock prices, fundamentally, two different approaches stand out.

Traditional evaluation approaches rely on the disciplines of economics and finance, utilizing fundamental and technical analysis methodologies. On one hand, fundamental analysis focuses on the intrinsic value of stocks and qualitatively assesses external factors such as interest rates, exchange rates, inflation rates, sector-specific policies, the financial conditions of companies listed on the stock exchange, and international market conditions. This includes economic and political dynamics as well.

On the other hand, technical analysis primarily focuses on trends in stock prices, trading volumes, and the psychological expectations of investors. This methodology is particularly geared towards analyzing the movements of individual stocks or overall market indices using price charts and various other tools. Today, traditional fundamental and technical analyses remain frequently favored methods among many institutions and individual investors(Lu et al., 2021); (Vijh et al., 2020).

In the academic literature, there are numerous indicators, ratios, oscillators, and metrics that can guide investors in forecasting stock prices. These techniques are based on the proposition of technical analysis, which assumes that past price movements will continue into the future. The assumption-based nature of these methods provides an explanation for why predictions associated with these indicators do not

E-mail address: taneruckan@yyu.edu.tr.

https://doi.org/10.1016/j.jksuci.2024.102162
Received 11 June 2024; Received in revised form 13 August 2024; Accepted 14 August 2024
Available online 27 August 2024

always yield statistically expected results. In this context, despite expectations formed through scientific methods, the existence of underlying assumptions in these approaches should be considered. Therefore, it is important to keep in mind that there can always be potential exceptions in every situation.(Akkaynak, 2023).

Researchers have extensively employed various statistical and machine learning methods for the forward prediction of stock prices. It has been observed that significant successful results are reported not only from models like ARIMA and SARIMA but also from other machine learning algorithms (Kaushik, 2020); (Ahmadpour et al., 2023).

Recently, artificial intelligence (AI) has begun to be increasingly integrated in the field of finance, especially in stock markets. Artificial intelligence algorithms have the ability to analyze large data sets and make predictions or decisions based on these analyses. These capabilities can be valuable in predicting stock prices, identifying market trends, and making investment decisions. For example, AI algorithms can predict future performance using a variety of data, such as companies' financial health, historical share prices, and stock market trends. Artificial intelligence can also identify stock trading opportunities by monitoring market conditions in real time. All in all, artificial intelligence has the potential to significantly improve the efficiency and accuracy of stock market analysis and decision-making, which can deliver better investment outcomes to investors(Gülmez, 2023).

In this article, 4 different models in which PCA and Deep learning models are used together are proposed and compared in detail in order to predict the prices of stocks for the next day. When estimating stock prices, it is necessary to make future predictions by taking past price movements into account. For this reason, deep learning models such as LSTM, where past information is kept in memory, are used in detail in this article. More than one indicator is used within the scope of technical analysis to predict stock prices. These indicators formulate past stock price movements in different ways and produce signals in the form of buy, sell and hold. In this study, the 10 most commonly used different indicators are analyzed with PCA (Principal component analysis) and the most effective ones are selected. The obtained effective indicators and stock price values are given as input values to the proposed deep learning models and the next day value is predicted.

The key contributions of the study are;

- *Integration of PCA and Deep Learning Models:* This study aims to increase the accuracy of price predictions in the Turkish stock market by investigating the integration of PCA and various deep learning models (LSTM, CNN, BiLSTM, GRU).
- *Selection of Effective Indicators:* In the study, the most effective technical indicators to be used in stock price prediction were selected and analyzed using the PCA method.
- *Performance Evaluation:* The performance of the proposed models was evaluated with Mean Square Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and R^2 metrics.
- *Superior Model Performance:* Among the tested models, the PCA-LSTM-CNN model showed superior performance especially in short-term price predictions and emphasized the benefits of integrating the indicators selected with PCA with deep learning models.
- Multiple hybrid deep learning models were tested and the most suitable model was determined.
- Separate tests were conducted with 30 and 60-day lag values to find the most accurate period for investing.

These contributions aim to provide investors and analysts with valuable tools to identify market opportunities and effectively manage risks, creating a solid framework for financial market analysis.

2. Related work

In the literature on stock analysis, there are many studies using both statistical-based methods and machine learning and deep learning models. Some recent studies in this field are detailed below and compared in Table 1.

(Gülmez, 2023) Related work presents a new deep LSTM network optimized with the ARO algorithm. The model is named LSTM-ARO and aims to make stock market price predictions. The data set used in the research covers price data of 30 different stocks of the DJIA index from 2018 to 2022. The duration of the data is determined as five years. For analysis, the data was reformatted to cover the previous 20 days to predict the next day's price. The developed model was evaluated by comparing it with LSTM1D, LSTM2D, LSTM3D, ANN and LSTM-GA models. The results show that the LSTM-ARO model exhibits superior performance and provides effective results compared to other models.

(Ma et al., 2023) In this study, researchers proposed a new method called Multi-source Aggregated Classification (MAC) to predict stock price movements. The MAC method combines numerical features of targeted stocks, market-based news sentiment, and news sentiment related to the specific stocks. To more effectively reflect real market sentiment from news sources, a model that extracts features from news in a way that aligns with actual stock price movements has been pretrained. Additionally, MAC utilizes a Graph Convolutional Network (GCN) to determine the impact of news on the targeted stocks. Experimental results demonstrate that MAC significantly enhances the accuracy of stock price movement predictions.

Table 1
Comparative analysis of related studies.

Study	Period	Dataset	Methodology/ models used
J. Li, Y. Liu, H. Gong, and X.Huang (2024)	2010–2019	SSEC, CSI500, SZSE,	Boruta,SVR,BSO
J. Cheng, S. Tiwari, D. Khaled, M. Mahendru, and U. Shahzad(2024)	2017–2022	Bitcoin(BTC)	Fb-Prophet, ML, SARIMA,LSTM
B. Gülmez(2023)	2018–2023	DJIA	ARO-LSTM,GA
W. Lu, J. Li, J. Wang, and L. Qin(2023)	2018–2021	Chinese stock market	NLP,Sentiment Analysis,BILSTM
S. Albahli, T. Nazir, M. Nawaz, and A. Irtaza(2023)	2013–2023	SP-500	DenseNet-41, Autoencoder,STI
M. Ali, D. M. Khan, H. M. Alshanbari, and A. A. A. H. El-Bagoury(2023)	2015–2022	KSE-100	EMD-LSTM
I. U. Armagan (2023)	1996–2023	BIST100 (XBANKS)	Fb-Prophet, ARIMA,CNN
G. Il Kim and B. Jang (2023)	2012–2021	Petroleum Oil Price	CNN-LSTM,CNN-GRU,Skip-Connection
F. Yang, J. Chen, and Y. Liu(2023)	2016–2020	–-	PSO,AI, Cyclic neural network
K. Yadav, M. Yadav, and S. Saini(2021)	2020	NYSE	FastRNNs,CNN, BiLSTM
H. Zhang(2018)	2013–2018	Chinese stock market	PCA,BP NN
(Zhang et al., 2024) Zhaofeng Z., Banghao C., Shengxin Z.*, Nicolas L.(2024)	2010–2019	Chinese stock market	Transformers, Sentiment Analysis
(Costa and Machado, 2023)Lorenzo D. C., Alexei M. C.M (2023)	2018–2023	Ibovespa index	Transformers, ARIMA,LSTM
(Wang, 2023) Shuzhen W.(2023)	2012–2023	A-share Index, Shanghai Composite Index, Shenzhen Component Index, CSI 300	Transformer, BiLSTM, TCN
(Sarıkoç and Celik, 2024)Sarıkoç, M., & Celik, M. (2024)	2000–2017	S&P500	PCA, ICA, LSTM

(Li et al., 2023) In their approach to improve model accuracy, the authors initially applied an adaptive empirical modal decomposition method to the primary data. Subsequently, they filtered technical indicator data using the Boruta method and enhanced the selected features with an adaptive noise reduction technique. In the final stage, they employed support vector regression (SVR) integrated with a brainstorm optimization algorithm (BSO) for effective data processing and prediction of target variables. The results obtained are reported to be promising.

(Cheng et al., 2023) This research conducted an empirical analysis on financial time series and machine learning techniques using Long Short-Term Memory (LSTM), Seasonal Autoregressive Integrated Moving Average (SARIMA), and Facebook Prophet models to predict Bitcoin prices and Garman-Klass (GK) volatility. The analysis results indicate that the LSTM model provides a noticeable performance improvement over the SARIMA and Facebook Prophet models in terms of Mean Squared Error (MSE) and Mean Absolute Error (MAE).

(Albahli et al., 2023) In this study, researchers used Stock Technical Indicators (STIs) to make final price predictions for products. To reduce highly correlated data, STIs were initially input into an autoencoder. The processed STIs and financial data were then fed into a DenseNet-41 model. The results have been used to predict short, medium, and long-term closing costs. The findings demonstrate that the model provides buy, sell, or hold signals to investors and outperforms existing methods.

(Ali et al., 2023) This study proposes a new hybrid model based on an enhanced version of Empirical Mode Decomposition (EMD) and a deep learning technique known as Long Short-Term Memory (LSTM) to improve the accuracy of complex stock market predictions. The forecast performance of the proposed hybrid Akima-EMD-LSTM model was evaluated using the KSE-100 index on the Pakistan Stock Exchange. The results reveal that the proposed model outperforms all other examined models, proving to be an effective tool for predicting financial time series.

(Armagan, 2023) In this study, traditional Autoregressive Integrated Moving Average (ARIMA) Model along with two artificial intelligence-based deep learning models, the Facebook Prophet Model (FPM), and Convolutional Neural Networks Model (CNNM), were used. Analysis results show that CNNM exhibited superior performance compared to other models.

(Il Kim and Jang, 2023) This research compared both univariate and multivariate methods to make more accurate and explanatory predictions. A new model inspired by DenseNet architecture was proposed to provide a more detailed and accurate analysis of oil price movements, enhanced with convolutional neural networks (CNN) and Long Short-Term Memory (LSTM) using skip connections. The results show that the proposed model performs better than traditional models.

(Yang et al., 2023) In this study, an advanced PSO algorithm was used to develop a neural network-based prediction model to enhance the accuracy of stock price predictions. The research results demonstrate that the developed system has a practical impact.

(Yadav et al., 2022) The authors presented two innovative models for different application scenarios. The first is based on Fast Recurrent Neural Networks (Fast RNN), which was used for the first time in this study for stock price predictions. The second model is a hybrid deep learning model that combines the strengths of Fast RNNs, Convolutional Neural Networks, and Bidirectional Long Short-Term Memory models to predict sudden changes in a company's stock prices. The results indicate that the proposed model performs better than traditional models.

3. Material and method

The proposed method for predicting future values of stock prices fundamentally consists of three stages. In the first stage, historical data from the past 10 years for five stocks listed on Borsa Istanbul are acquired via Yahoo Finance(Yahoo Finance, 2024). Subsequently, these data are scaled, and values for the ten most commonly used indicators in technical analysis are calculated. In the second part of the study, the indicator values obtained are analyzed using Principal Component Analysis (PCA) to identify the most influential indicators on stock prices. In the final stage, the features obtained and the closing prices of the stocks are used to make price predictions for the stocks using four different proposed deep learning models. Among the models tested in our study, the PCA-LSTM-CNN model is recommended due to its superior performance. The flowchart of the recommended model is shown in Fig. 5.

3.1. Data preprocessing

This section describes the dataset used, and the preprocessing performed on the dataset prior to the application. The historical data used for the training and testing processes of the proposed model were obtained from the Yahoo Finance website, covering the period from January 1, 2014, to January 1, 2024. The constructed dataset comprises seven columns: Date, Open, High, Close, Adj Close, and Volume. Detailed information about these fields is provided in Table 2.

As no missing values were observed in the collected data, the entirety of the dataset was utilized. To ensure the optimal use of the proposed models and to facilitate their correct training, the data was subjected to a normalization process. During this normalization stage, the data was scaled to a range between 0 and 1 using a min–max scaler, employing the equation provided in equation (1) (Ratchagit and Xu, 2022); (Zhang, 2018).

$$x_p = \frac{x_p - x_{min}}{x_{max} - x_{min}} \tag{1}$$

3.2. Dataset

In this study, it is aimed to predict future price values using financial time series data. The data used for this purpose was obtained from the Yahoo Finance website via the yfinance library. As seen in Fig. 1, the obtained data is divided into training, validation, and test data. While 80 % of the data is allocated for the training dataset, 20 % for validation, and 20 % for testing.

Within the scope of the study, ten indicators most frequently used in the field of technical analysis are used. The indicator values obtained as a result of mathematical calculations of these indicators are shown in Fig. 2.

The created dataset consists of 10-year data between 2014 and 2024. In the first stage, it consists of 2317 rows and 16 columns with the data and indicator values of the stock. After the feature reduction with PCA, a dataset consisting of 2317 rows and 3 PCs is created. In Fig. 3, the relationship between the created PC values and the real price of the stock is shown graphically.

3.3. Stock technical indicators (STIs)

Stock market indicators are mathematical calculations or statistical metrics that assist investors and analysts in analyzing the price movements of stocks. These indicators utilize historical price and volume data

Table 2
Fields and descriptions in the data set.

Name	Description
Date	Defines the date information for the relevant day.
Open	Defines the opening price for the relevant day.
High	Defines the highest price on the relevant day.
Low	Defines the lowest price on the relevant day.
Close	Defines the closing price for the relevant day.
Adj. Close	Adjusted Closing Price is the adjusted version of stock prices in line with corporate movements such as dividends and stock splits.
Volume	indicates the total transaction volume

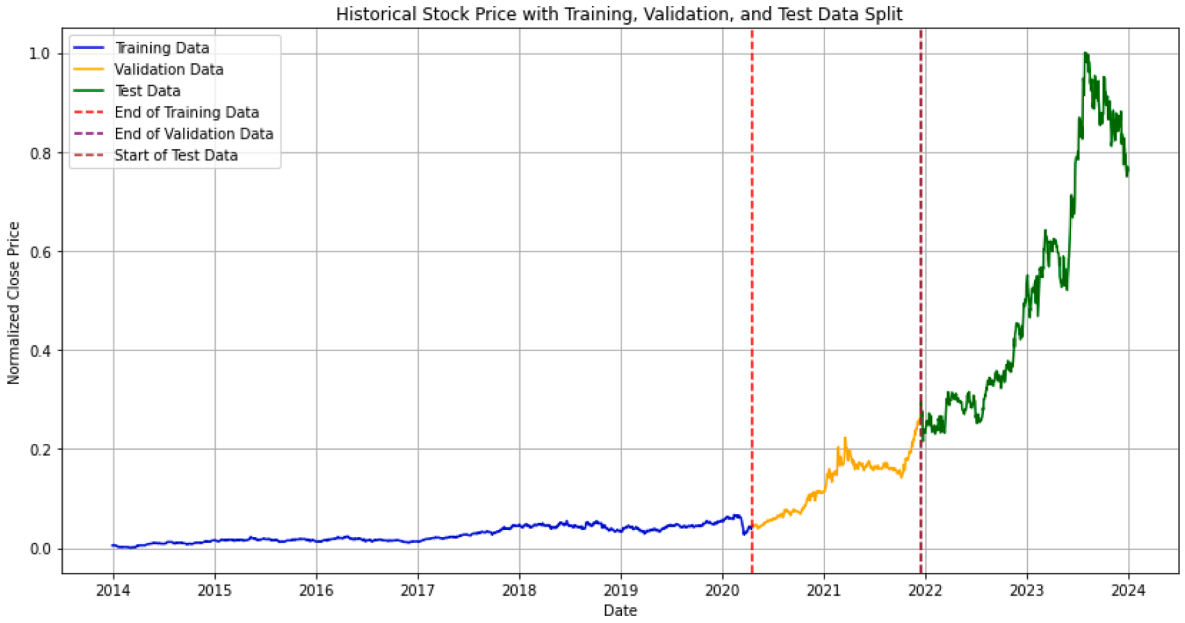

Fig. 1. Historical stock price with training, validation and test data split.

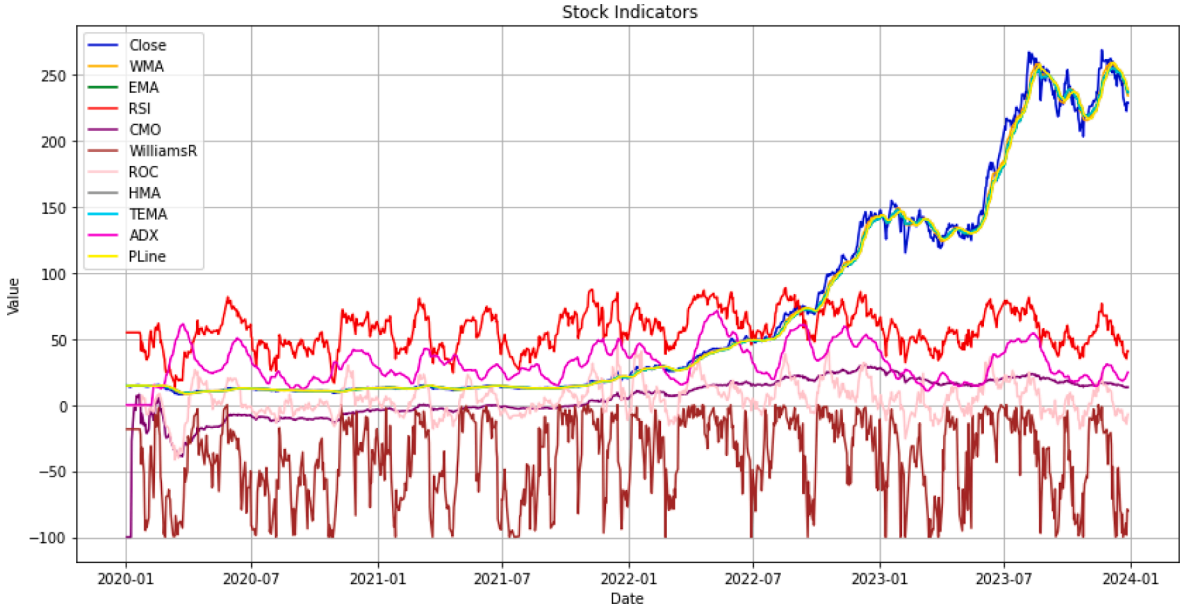

Fig. 2. Graphical representation of the technical indicator.

of stocks to provide insights into market trends and potential price changes(Akar and Gur); (Kabir Ahmed et al., 2019). The technical indicators used within the scope of this study are listed below.

3.3.1. *Weighted moving average (WMA)*

The Weighted Moving Average is a type of moving average calculated by assigning different weights to data points. This method is particularly used to forecast trends in data changes over a specified period(Rifai, 2024). The calculation of the Weighted Moving Average is as specified in equation (2).

$$WMA = \frac{Price_1 + Price_2 x(n-1) + \cdots Price_n}{\frac{nx(n+1)}{2}} \tag{2}$$

n = Time Period

3.3.2. *Exponential Moving Average (EMA)*

The Exponential Moving Average (EMA) is a technical indicator frequently used in financial analyses. EMA responds more rapidly than previous moving averages by giving greater importance to the most recent data. This characteristic allows the EMA to reflect changes in price trends more quickly, thereby offering investors the opportunity to adapt more swiftly to current market conditions(MRG and Panchal, 2021). The calculation of the Exponential Moving Average is as specified in equation (3).

$$EMA_{today} = \left(Price_{today} x SF\right) + \left(EMA_{Previousday} x(1-SF)\right) \tag{3}$$

where,

$Price_{today}$: Today's price.

$EMA_{Previousday}$: EMA value of the previous day.

SF: Smoothing factor.

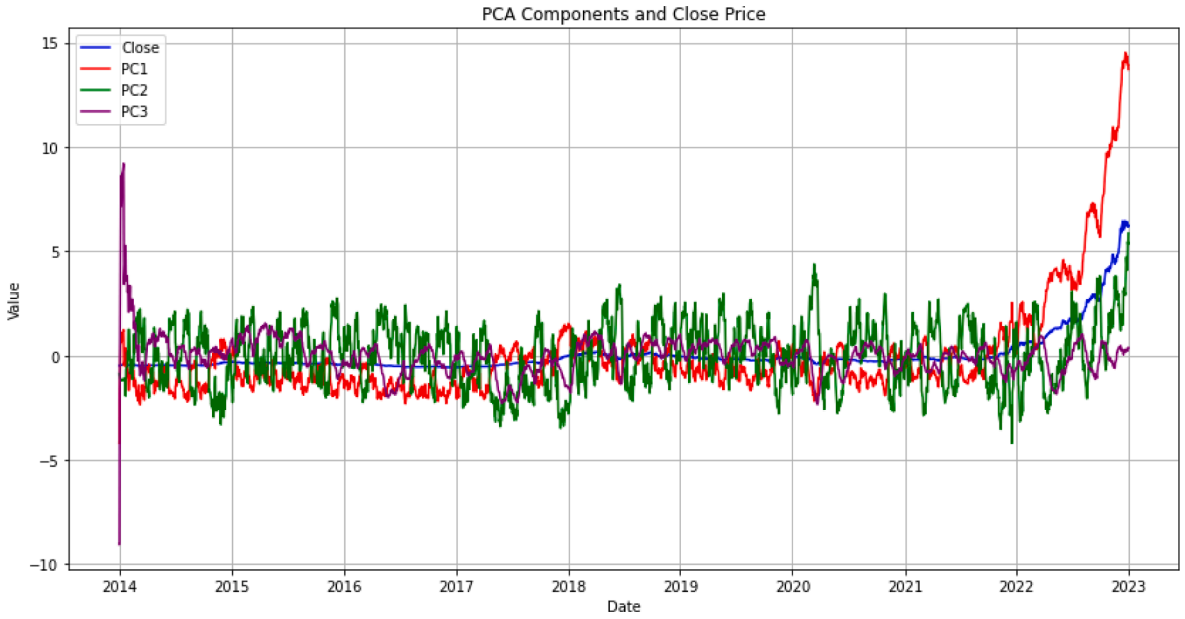

Fig. 3. PC values and the real price of the stock.

N: Represents the period of the moving average.

3.3.3. Relative strength Index (RSI)

The Relative Strength Index (RSI) is a technical analysis tool used to measure the momentum of price movements in financial markets. Developed in 1978 by J. Welles Wilder, the RSI helps identify overbought or oversold conditions of a stock or other financial asset(Indah and Mahyuni, 2022). The calculation of the RSI is as specified in equation (4).

$$RS = \frac{AverageGain}{AverageLoss}$$

$$RSI = 100 - \left(\frac{100}{1 + RS}\right) \tag{4}$$

3.3.4. Chande momentum oscillator (CMO)

The Chande Momentum Oscillator (CMO) is grounded in a methodology distinct from other momentum indicators such as the Relative Strength Index (RSI) and the Stochastic Oscillator (KDJ). Developed by Tushar Chande, this indicator incorporates data from both rising and falling days in its calculations(Hermitian and Quantum, 2022). This characteristic enables the CMO to provide a more comprehensive analysis of market momentum, thereby offering investors and market analysts enhanced capabilities to effectively identify both overbought and oversold conditions. The calculation of the CMO is as specified in equation (5).

$$CMO = 100x\left(\frac{SumUp - SumDown}{SumUp + SumDown}\right) \tag{5}$$

where;

SumUp: Represents the sum of positive changes, namely price increases.

SumDown : Represents the sum of the absolute values of negative changes, namely price decreases.

3.3.5. Williams percent range (WILLR)

Williams Percent Range (%R), developed by Larry Williams, is a momentum indicator used to detect overbought or oversold conditions in the market. This indicator evaluates an asset's closing price in relation to its highest and lowest prices within a recent period. Typically, Williams %R values range between −100 and 0. (Pandya and Jaliya, 2021). The calculation of Williams %R is as specified in equation (6).

$$Williams\%R = \left(\frac{HighestPrice - ClosingPrice}{HighestPrice - LowestPrice}\right)x - 100 \tag{6}$$

3.3.6. Rate of change (ROC)

Rate of Change (ROC) is one of the oldest and simplest technical indicators. It simply measures the discrete return of the current market price relative to an older market price(Deszi and Scarlat, 2013, 2013). The calculation of ROC is as specified in equation (7).

$$ROC = \left(\frac{CurrentPrice - PriceNdaysago}{PriceNdaysago}\right)x100 \tag{7}$$

3.3.7. Hull Moving Average (HMA)

The Hull Moving Average (HMA) is a type of composite moving average that consists of a combination of Weighted Moving Averages (WMA) calculated over different time intervals. (Raudys, 2014). The calculation of HMA is as specified in equation (8).

$$HMA_n = WMA_{\sqrt{n}}\left(2xWMA_{\frac{n}{2}} - WMA_n\right) \tag{8}$$

where:

- n represents the period.
- WMA_n represents the Weighted Moving Average for n- periods
- $WMA_{\frac{n}{2}}$ represents the Weighted Moving Average for $\frac{n}{2}$ periods
- $WMA_{\sqrt{n}}$ is represents the Weighted Moving Average for $\sqrt{n}$ periods.

3.3.8. Triple exponential Moving Average (TEMA)

The triple moving average trading strategy involves using the Triple Exponential Moving Average (TEMA) or Triple Simple Moving Average (TSMA) to follow market trends and identify trading opportunities. This strategy employs a combination of moving averages over three different periods, enabling a broader perspective analysis of market movements. (Walugembe and Stoica, 2022). The calculation of TEMA is as specified in equation (9).

$$TEMA = 3xEMA_1 - 3xEMA_2 + EMA_3 \tag{9}$$

- EMA_1 It is a first-order exponential moving average.
- EMA_2 It is a second-order exponential moving average (EMA of EMA_1).
- EMA_3 It is the third- order exponential moving average (EMA_2′s EMA)

3.3.9. Average Directional index (ADX)

The Average Directional Index (ADX) indicator is a technical analysis tool used to measure the strength or weakness of trends in the market. An increasing ADX value does not indicate that prices are moving in a certain direction; however, it suggests that the current trend is strong. (Journals, 2019).

3.3.10. Psychological line (PLine)

Pline is an important indicator calculated by the ratio of the total number of rising days to the total days and used to compare sales and purchasing power. (Albahli et al., 2023). The calculation of Pline is as specified in equation (10).

$$Pline = \left(\frac{D_{up}}{D_{total}}\right) x100 \tag{10}$$

where;

D_{up}: The number of rising days in the selected period.

D_{total}: The total number of days in the selected period.

3.4. Principal Component analysis (PCA)

Within the scope of this study, 10 technical indicator values are calculated together with the closing information of the stock. Due to the large number of parameters in the model to be used, PCA, one of the feature reduction methods, was used. PCA is a technique used to reduce the size of a data set containing multiple variables This methodology functions by examining the covariance between variables and moves the original data into a coordinate system that is reordered by the amount of variance. PCA uses a mathematical process that converts correlated variables into linearly independent variables called 'principal components'. In the dimensionality reduction process, the first few components containing the highest variance are generally preferred. Each subsequent component is chosen to explain as much of the available variance as possible. PCA is especially useful in cases with multiple highly correlated dimensions; this technique can reduce data redundancy by focusing on a few independent components that can explain a significant amount of variance (Zhang, 2018); (Wen et al., 2020).

There is a high degree of correlation between the stock technical indicators used in our study. This situation can negatively affect the performance of the model because redundant information can cause the model to overfit. PCA eliminates this highly correlated data and allows us to obtain more independent and meaningful features from the principal components. This process reduces the complexity of the model and increases the performance of the model with fewer but more informative features. PCA reduces the redundant information in the data set, allowing the model to work more efficiently. Especially in high-dimensional data sets, redundant information can reduce the generalization ability of the model. With PCA, this redundant information is eliminated, allowing the model to make more accurate predictions. PCA reduces the data to more meaningful components and thus improves the performance of the model. This helps the model to give better results, especially in test data sets. How the indicators selected with PCA increase the prediction accuracy of the model is supported by the experiments and the results obtained.

Although deep learning models are capable of processing large data sets, the use of PCA increases computational efficiency. Training the model with fewer features both reduces computational time and allows the model to respond faster. This is a significant advantage, especially when working with large data sets. PCA allows the model to use fewer computational resources during the training process. This is a significant advantage, especially in systems with limited computational power.

Each of the basic components obtained as a result of PCA analysis explains a certain percentage of the variance in the data set. It can be easily understood by looking at Fig. 4 that the first three basic components (PC) explain the total variance to a large extent. Which technical indicators carry more information about these components and how they represent certain trends and patterns in stock price movements are analyzed in Table 3**,** and it is observed that the Williams R, ADX, and HMA indicators contribute more to the total variance. The first few basic components explain a large portion of the total variance, and these components were used to increase the model's predictive performance. These components were derived from the indicators that provide the highest contribution to stock price predictions. The effects of these indicators on stock prices were examined in detail.

3.5. Models

Within the scope of this article, four different models were created using different deep learning methods. The parameters given to the input layer in the created models are first selected after going through the PCA step. The selected indicator value is given to the input layer along with the closing price of the share price. Information about the created models and the proposed model is given in detail in Table 4.

The hyperparameters used in the learning process in the proposed deep learning model and other models were chosen the same to provide a fair approach in the model comparison process. These hyperparameters are presented in detail in Table 5.

4. Proposed model

In this study, a hybrid deep learning architecture is proposed for stock price prediction by utilizing both temporal and feature-based learning from historical stock data. Daily closing prices of stocks and three principal component analysis (PCA) selected indicator values necessary to capture fundamental trends and market movements are used as model inputs.

4.1. Model architecture

A detailed representation of the hybrid model proposed within the scope of the study is shown in Fig. 5.

Input Layer: The input layer receives normalized stock data, which is free from scale differences and focuses on extracting meaningful patterns.

Long Short Term Memory (LSTM) Layer*:* LSTM is a type of recurrent neural network that can comprehend long-term relationships and solve the short-term commitment problem(Shohan et al., 2022). By regulating the flow of information through gate mechanisms, it has the ability to successfully capture patterns in extended time series, unlike standard RNNs. LSTM networks are suitable for time series forecasting due to their ability to remember information over long periods of time. This layer helps the model understand long-term dependencies that affect stock prices. LSTM networks can remember this information over a longer period of time by using memory cells, input gates, output gates, and forget gates. LSTM (Long Short-Term Memory) units define how data will be processed through specific mathematical equations. There are four main components in an LSTM cell: forget gate (f), input gate (i), output gate (o), and cell state (C)(Tsilingeridis et al., 2023). The basic equations defining these components are given below in equations (11)–(16).

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_{f)} \tag{11}$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{12}$$

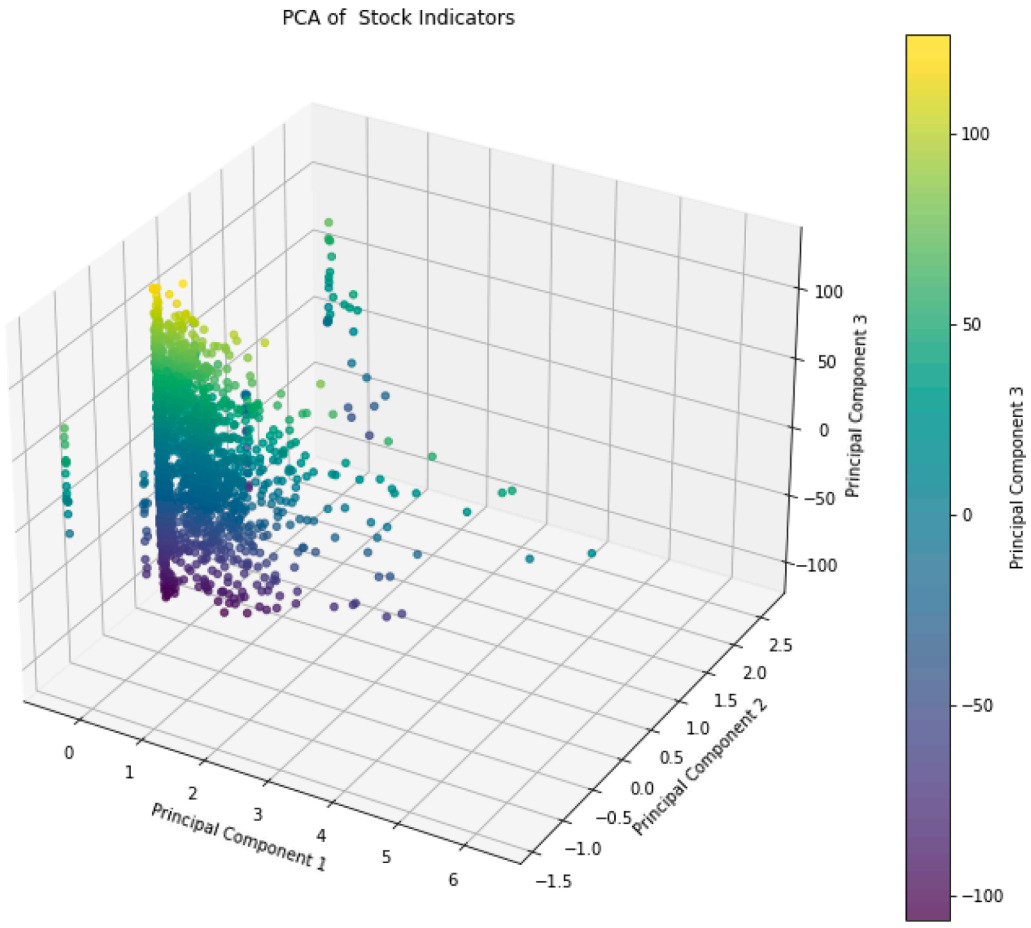

Fig. 4. PCA variance distribution graph.

Table 3
Contribution of components to variance.

Indicators	Contributions of Component		
	PCA 1	PCA 2	PCA 3
WMA	6.37 e-08	3.92 e-07	0.03
EMA	6.57 e-08	3.87 e-07	0.03
RSI	6.25 e-08	3.98 e-07	0.03
CMO	6.43 e-08	3.94 e-07	0.03
WilliamsR	**0.99**	7.99 e-06	1.20e-07
ROC	6.20 e-08	3.37 e-07	0.042
HMA	6.12 e-08	3.49e-07	**0.043**
TEMA	5.06 e-08	1.51 e-06	−0.34
ADX	1.01 e-07	**3.03 e-06**	−0.69
PLine	6.55e-08	2.32 e-06	−0.56

$$\widetilde{C}_t = \tanh(W_C{*}[h_{t-1}, x_t] + b_C) \tag{13}$$

$$C_t = f_t{*}C_{t-1} + i_t{*}\widetilde{C}_t) \tag{14}$$

$$o_t = \sigma(W_o{*}[h_{t-1}, x_t] + b_{0)} \tag{15}$$

$$h_t = o_t{*}\tanh(C_t) \tag{16}$$

The internal structure of an LSTM cell is as shown in the Fig. 6.

Convolutional Neural Network (CNN): A one-dimensional CNN model has a convolutional hidden layer operating on a 1-dimensional array (Gamboa, 2017). CNN has three basic components: convolution layers, pooling layers, and fully connected output layers. The main elements of this architecture are convolution and pooling operations. Convolution layers enable the detection of local features, while pooling layers help reduce the risk of overlearning and reduce the number of parameters (Hu et al., 2021; Wu et al., 2023; Pabuccu and Barbu, 2023). he first layer of our architecture is the 1D Convolution layer. This layer applies a series of filters to the input data to extract high-level features from stock indicators. By gliding over input data with a specified kernel

Table 4
The layers and parameter values of the deep learning models.

Model	Layer	Parameter
Model_1	Input	4
	Conv1D	64
	MaxPooling1D	2
	LSTM	50
	GRU	50
	Dense	1
Model_2	Input	4
	Conv1D	50
	BiLSTM	50
	Dropout	0.5
	Dense	25
	Dense	1
Model_3	Input	4
	LSTM	50
	LSTM	50
	Dense	1
Proposed Model	Input	4
	LSTM	50
	Conv1D	50
	MaxPooling1D	2
	Flatten	
	Dense	25
	Dense	1

Table 5
Hyperparameters.

Hyperparameter	Values
Optimizer algorithm	Adam
Learning rate	0.01
Activation Function	Relu
Padding	same
Lag values (Day)	30,60
Train set size	%80
Test set size	%20
Los function	mean_squared_error
Validation split	0.2

size, it captures local dependencies and patterns that are critical to understanding short-term trends. The internal structure of an CNN cell is as shown in the Fig. 7.

4.2. Evaluation measures

The effectiveness of the proposed model within the scope of this study was verified by extensive backtests based on historical data and compared with other models. Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and R Squared (R2) measurements, which are frequently used in the literature, are used to evaluate performance measurements.

MSE: Mean Squared Error (MSE) is the average of the squares of the differences between predicted values and actual values. The calculation of MSE is as specified in equation (17).

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 \tag{17}$$

Here y_i is the actual values, $\widehat{y}_i$ is the predicted values and n is the number of samples.

MAPE: Mean Absolute Percentage Error is a statistic that measures how accurate model forecasts are relative to observed values, and is often used in various forecasting problems such as time series forecasts.. The calculation of MAPE is as specified in equation (18).

$$\left(\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\widehat{y}_i}{y_i}\right|\right)*100 \tag{18}$$

MAE: Mean Absolute Error (MAE) is the average of the absolute differences between predicted values and actual values. MAE gives a direct idea of the scale of errors and is less affected by outliers. The calculation of MAE is as specified in equation (19).

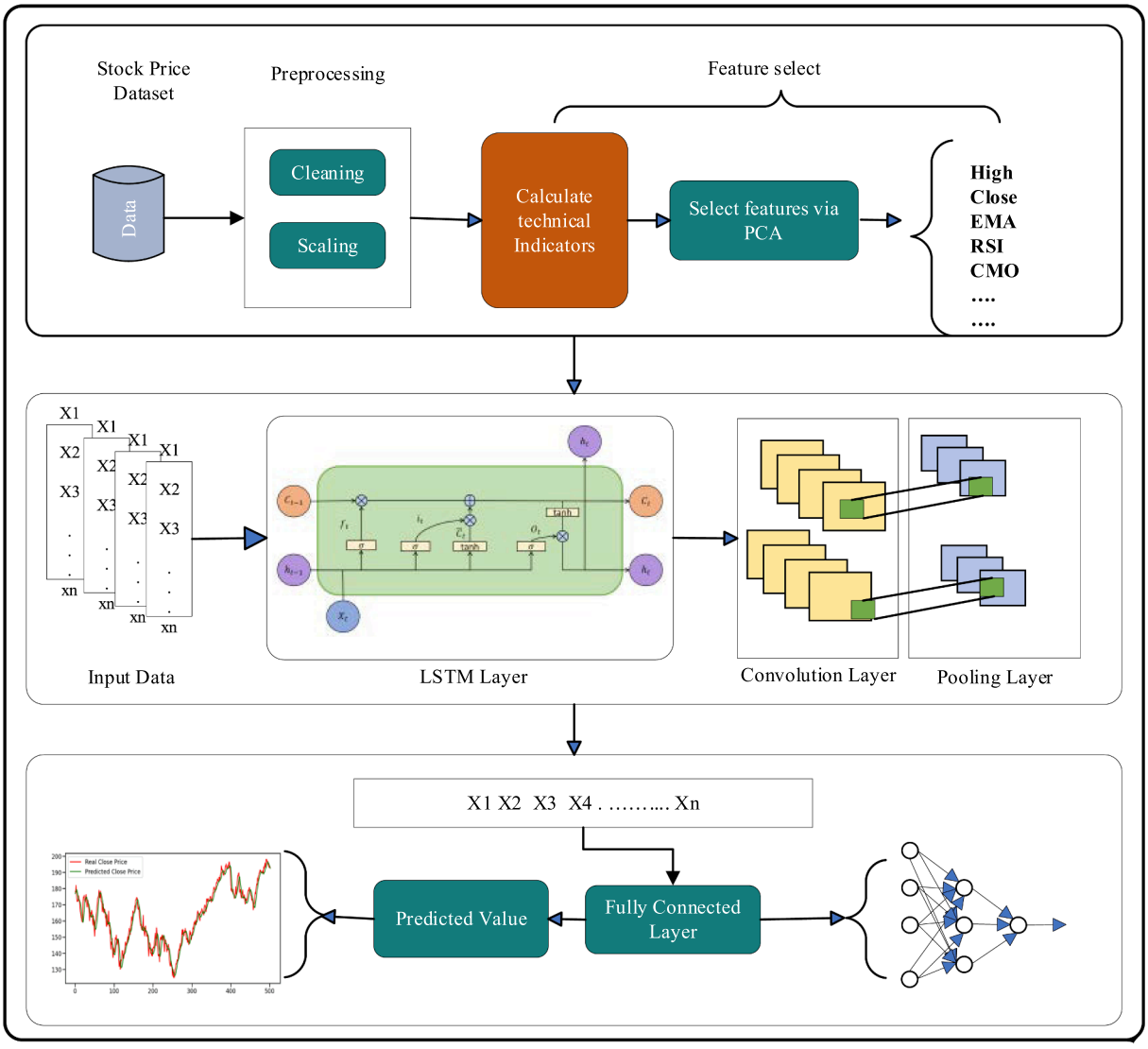

Fig. 5. Proposed model architecture.

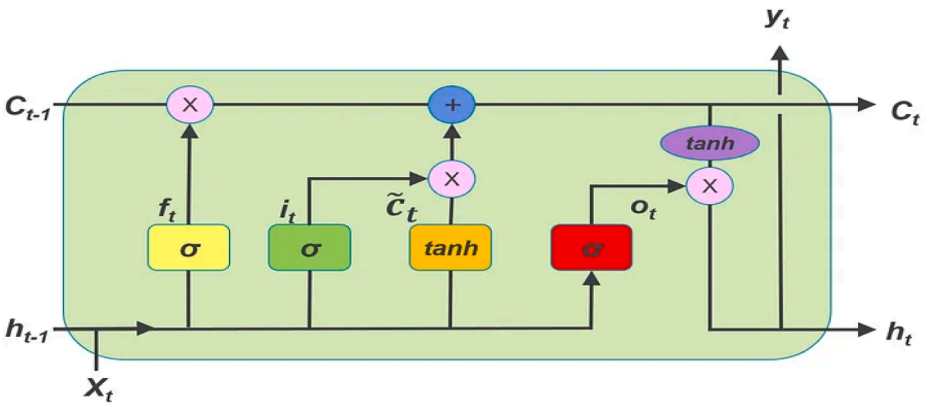

Fig. 6. Internal structure of an LSTM cell.

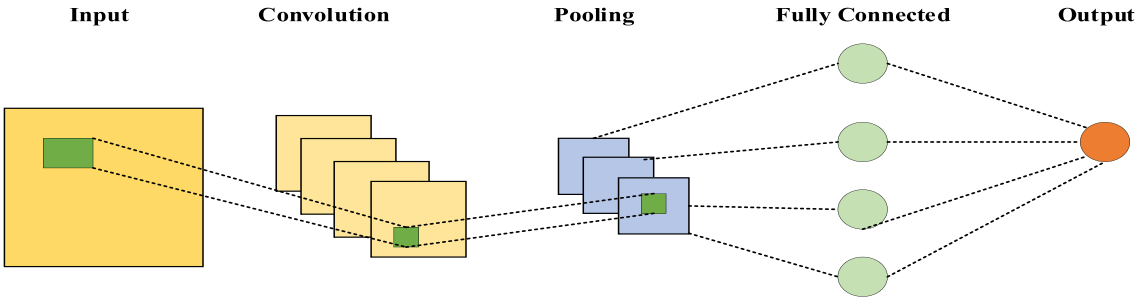

Fig. 7. Fully convolutional neural network architecture.

$$\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \tag{19}$$

R^2: It is a statistical measure that measures how well a regression model fits the data. Their values range from 0 to 1; The closer it is to 1, the better the model fits the data. The calculation of R^2 is as specified in equation (20).

$$R^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{20}$$

Here y_i refers to the actual values, $\hat{y}_i$ refers to the values predicted by the model, $\bar{y}$ refers to the arithmetic mean of the actual values and n refers to the number of observations.

5. Result

In this study, 4 separate models were created in which PCA and Deep learning models were integrated with each other for the prediction of stock prices and the most effective model was proposed. The proposed models are named model_1, model_2 and model_3, respectively. In order to test the accuracy of the models, stock prices of 5 well-established companies in the BIST100 index of the Istanbul Stock Exchange were examined. When the results obtained are examined, the proposed model (PCA-CNN-LSTM) produced very successful results compared to other models. Graphs of the results are given in Fig. 8 and Fig. 9. In Fig. 8, forecast results with 30-day lag values for all stocks are presented, while in Fig. 9, the results with 60-day lag values are shown graphically.

According to the MSE values in Table 6, it is observed that the proposed model yields the lowest values in 30 and 60-day lagged data for each stock compared to the other three models. Particularly, the extremely low MSE values of 0.0005 obtained in the 30-day data for ASELS, TUPRS, and SISE stocks indicate that the proposed model predicts the price movements of these stocks more accurately than other models. Furthermore, upon examining the 60-day lag values, it is evident that the proposed model excels in many instances and delivers similar results in others. Among the other models, model_2 generally attains lower MSE values than Model_1 and Model_3, signifying its second-best performance.

Table 7 displays the MAPE values, which measure the percentage deviation of predictions from actual values. The proposed model exhibited the lowest deviation rates for all stocks, as evidenced by the MAPE results. For instance, the proposed model's MAPE value of 3.46 % for ASELS, 3.67 % for TUPRS, 4.9 % for THYAO, 2.8 % for SISE, and 5.5 % for FROTO in 30-day data was significantly lower than the others. Similar performance was observed for 60-day delay values. Model_2 generally outperformed Model_1 and Model_3, providing lower percentage errors. These findings suggest that the proposed model's predictions are more sensitive to market changes and percentage-wise closer to the real values.

Table 8 shows that the proposed model exhibits the lowest MAE values across all stocks, suggesting that its prediction errors have a smaller absolute magnitude compared to other models. Notably, for ASELS and TUPRS stocks, the proposed model's low MAE values (0.0165 and 0.0156, respectively) highlight its exceptional accuracy in forecasting the prices of these stocks. In comparison to Model_2, Model_1, and Model_3, the other models generally attained lower MAE values, indicating a more consistent performance.

Finally, when the R^2 values in Table 9 are examined, it is seen that the proposed model has higher coefficients than other models in all stocks, which indicates that the model is the model that best explains the variance in the data set. Values such as 0.990 for ASELS and 0.991 for TUPRS show that the proposed model almost perfectly reflects the price movements of these stocks and its predictions are highly reliable. Model_2 generally gave better results than Model_1 and Model_3 in R^2 values and stood out as the second most effective model.

In this study, various performance indicators such as MSE, MAE, MAPE, and R^2 were used to comprehensively evaluate the model

Fig. 8. Results from the 30-day lag values of the proposed model.

performance. These indicators allow us to evaluate the accuracy and generalization ability of the model from different perspectives. In order to determine the most appropriate performance indicator for predicting stock prices, it is important to consider both the accuracy of the model and the magnitude of the prediction errors. The study by (Chicco et al., 2021) provides detailed information on how performance indicators can be used in different situations. The study concluded that the R^2 indicator is more informative and reliable compared to other common indicators such as MSE, MAE, MAPE, and RMSE. In this context, the R-squared (R^2) indicator is considered the most appropriate for predicting stock prices. R^2 evaluates the explanatory power of the model and the accuracy of the predictions holistically. In addition, MSE should also be taken into account in cases where large errors are significant and need to be reduced.

6. Discussion

In this study, the effect of the integration of PCA and deep learning models on the forecasting performance of stock prices of five companies traded on Borsa Istanbul was investigated. Four different models were used in the study, and these models were trained with PCA-selected indicators and stock closing prices. Within the scope of the study, firstly, the technical indicators that are thought to be most related to the stock price and most effective in predicting future price formations were determined, and at the same time, it was aimed to eliminate the ones that are not very effective from the ten indicators used. Thus, effective features were selected by reducing the features. Afterward, four different deep learning models were created and the PCA-LSTM-CNN model that produced the best forecast result was proposed. With the proposed model, investors first made feature selection and reduction, then LSTM was used to recognize and learn past data patterns, and finally, local patterns and dependencies were captured with CNN and prediction was made. When the obtained results are taken into consideration, it is seen that it can offer suggestions with a very high success rate for investors to make future investments. The study also has some limitations. These limitations can be summarized as follows.

- The research is limited to stock data of only five companies traded on Borsa Istanbul. Therefore, the results obtained may not be directly generalizable to stocks in different market conditions and different geographical regions.
- The dataset used in the study covers a 10-year period between 2014 and 2024. The impact of economic, political, and social events within this time period has not been evaluated. Analyses conducted in different time periods may yield different results.

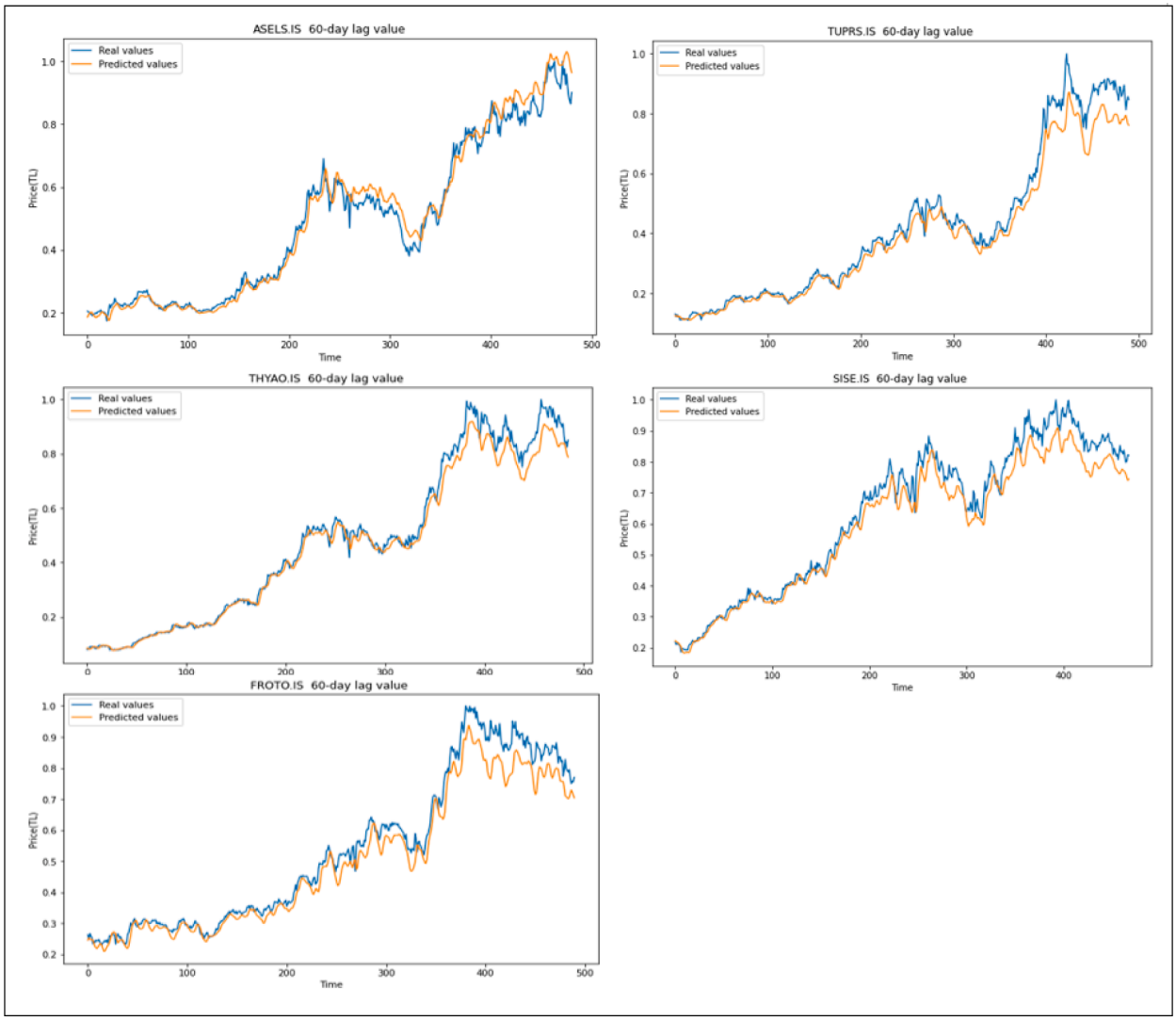

Fig. 9. Results from the 60-day lag values of the proposed model.

Table 6
Comparison of the models for MSE criteria.

MSE	30 days lag values				60 days lag values			
Ticker	**Model_1**	**Model_2**	**Model_3**	**Proposed Model**	**Model_1**	**Model_2**	**Model_3**	**Proposed Model**
ASELS	0.0060	0.0011	0.0040	**0.0005**	0.0059	0.0044	0.0032	**0.0012**
TUPRS	0.0068	0.0016	0.0017	**0.0005**	0.0079	0.0034	**0.0009**	0.0021
THYAO	0.0037	0.0023	0.0033	**0.0011**	0.0078	**0.0023**	0.0046	0.0028
SISE	0.0027	0.0032	0.0071	**0.0005**	0.0049	0.0044	0.0086	**0.0018**
FROTO	0.0047	0.0031	0.0043	**0.0014**	0.0071	0.0045	0.0074	**0.0021**

Table 7
Comparison of the models for MAPE criteria.

MAPE	30 days lag values				60 days lag values			
Ticker	**Model_1**	**Model_2**	**Model_3**	**Proposed Model**	**Model_1**	**Model_2**	**Model_3**	**Proposed Model**
ASELS	0.0882	0.0597	0.0741	**0.0346**	0.0850	0.0884	0.0737	**0.0432**
TUPRS	0.1251	0.0605	0.0551	**0.0367**	0.1292	0.0921	**0.0430**	0.0793
THYAO	0.0679	0.0557	0.0686	**0.0491**	0.1334	**0.0547**	0.0803	0.0643
SISE	0.0675	0.0646	0.0958	**0.0282**	0.0964	0.0867	0.112	**0.0547**
FROTO	0.0942	0.0690	0.0840	**0.0557**	0.1126	0.0783	0.1053	**0.0642**

Table 8
Comparison of the models for MAE criteria.

MAE	30 days lag values				60 days lag values			
Ticker	Model_1	Model_2	Model_3	Proposed Model	Model_1	Model_2	Model_3	Proposed Model
ASELS	0.0554	0.0272	0.0454	**0.0165**	0.0540	0.0507	0.0428	**0.0243**
TUPRS	0.0613	0.0283	0.0277	**0.0156**	0.0657	0.0433	**0.0201**	0.0348
THYAO	0.0416	0.0320	0.0405	**0.0241**	0.0701	**0.0323**	0.0485	0.0365
SISE	0.0441	0.0455	0.0693	**0.0176**	0.0619	0.0572	0.0787	**0.0350**
FROTO	0.0555	0.0424	0.0514	**0.0302**	0.0677	0.0499	0.0665	**0.0360**

Table 9
Comparison of the models for R2 criteria.

R2	30 days lag values				60 days lag values			
Ticker	Model_1	Model_2	Model_3	Proposed Model	Model_1	Model_2	Model_3	Proposed Model
ASELS	0.895	0.979	0.929	**0.990**	0.897	0.923	0.943	**0.978**
TUPRS	0.8907	0.973	0.972	**0.991**	0.872	0.943	**0.985**	0.965
THYAO	0.956	0.972	0.960	**0.986**	0.906	**0.971**	0.945	0.965
SISE	0.950	0.940	0.869	**0.989**	0.908	0.917	0.840	**0.964**
FROTO	0.917	0.945	0.924	**0.974**	0.877	0.921	0.871	**0.963**

- The hyperparameters of the deep learning models used are limited to certain values. Optimizing these parameters with different values may change the model performance.
- Economic and political factors affecting stock prices have not been taken into account in the study. The impact of such external factors may affect model performance and reduce prediction accuracy.
- The training and testing processes of deep learning models are time-consuming processes. Higher-performance hardware can be used to shorten these processes.

7. Conclusion

In this article, a hybrid model is proposed in which the most effective indicator values for stock analysis are used together with the price values of the stock. PCA was used to decide the most effective indicator among the technical indicators used. The deep learning network is trained and tested with the obtained indicator values and stock information. At this stage, 10-year data between 01.01.2014 and 01.01.2024 for companies with ASELS, TUPRS, THYAO, SISE and FROTO codes are obtained with the help of the yfinance website. 80 % of this information is divided as training data and 20 % as test data. Stock information is time series data, so it needs to be formatted in a certain format in order to predict the next day's values. At this stage, the reformatting process was carried out by taking into account the previous day values of 30 and 60 days. Within the scope of the study, experiments were carried out on four models named model_1 (PCA-CNN-LSTM-GRU), model_2 (PCA-CNN-BiLSTM), model_3 (PCA-LSTM) and the proposed model (PCA-LSTM-CNN) and as a result of detailed measurements, the most PCA-LSTM-CNN model, which offers good performance, is proposed. It has been observed that the proposed model gives the best performance for each stock at all measurement values, especially for 30-day lag periods. At the same time, it is seen that it provides very good results in the vast majority of 60-day delay values.

When the data obtained is examined, it is seen that the proposed model consistently exhibits superior performance in stock price predictions and generally offers more accurate and reliable predictions compared to other models. These findings show that the proposed model has significant potential for use in financial markets.

In order to further increase the performance and reliability of the proposed model, some limitations detailed in Section six can be eliminated. In particular, optimizing the hyperparameters of the proposed deep learning models or analyzing social and political reactions from the latest pre-trained NLP models and integrating them into the deep learning model are thought to be useful in increasing the performance of the system.

CRediT authorship contribution statement

Taner Uçkan: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahmadpour, A., Mirhashemi, S.H., Panahi, M., 2023. Comparative evaluation of classical and SARIMA-BL time series hybrid models in predicting monthly qualitative parameters of Maroon river. Appl. Water Sci. 13 (3), 1–10.

A. Akar and ¨O Ü Gur, "Technical Indicators and LSTM Prediction for Stock Prices.".

Akkaynak, B., 2023. A study on the comparison of technical indicators used in stock price prediction with the BAHP method. J. Life Econ. 10 (1), 1–15.

Albahli, S., Nazir, T., Nawaz, M., Irtaza, A., 2023. An improved DenseNet model for prediction of stock market using stock technical indicators. Expert Syst. Appl. 232 (June), 120903.

Ali, M., Khan, D.M., Alshanbari, H.M., El-Bagoury, A.A.A.H., 2023. Prediction of complex stock market data using an improved hybrid EMD-LSTM model. Appl. Sci. 13 (3), pp.

Armagan, I.U., 2023. Price prediction of the Borsa Istanbul banks index with traditional methods and artificial neural networks. Borsa Istanbul Rev. 23, S30–S39.

Cheng J, Tiwari S, Khaled, M. Mahendru, Shahzad U, "Forecasting Bitcoin prices using artificial intelligence: Combination of ML, SARIMA, and Facebook Prophet models," Technol. Forecast. Soc. Change, vol. 198, no. May 2023, p. 122938, 2024.

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, 1–24.

L. D. Costa and A. M. C. Machado, "Prediction of Stock Price Time Series using Transformers," pp. 85–95, 2023.

Deszi, D., Scarlat, E., 2013. The Performance of ROC on the BSE. Fi Ba 2, 373–379.

J. C. B. Gamboa, "Deep Learning for Time-Series Analysis," 2017.

Gülmez, B., 2023. Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm. Expert Syst. Appl. 227 (January), 120346.

Hermitian, C., Quantum, N., 2022. ConvLSTM coupled economics indicators quantitative trading decision model. Symmetry (basel).

Hu, Z., Zhao, Y., Khushi, M., 2021. A survey of forex and stock price prediction using deep learning. Appl. Syst. Innov. 4 (1), 1–30.

Il Kim, G., Jang, B., 2023. Petroleum price prediction with CNN-LSTM and CNN-GRU using skip-connection. Mathematics 11 (3), pp.

Indah, Y.R., Mahyuni, L.P., 2022. The accuracy of relative strength index (RSI) indicator in forecasting foreign exchange price movement. Inovbiz J. Inov. Bisnis 10 (1), 96.

Journals I.S., "Technical analysis indicators: pathway towards rewarding journey," vol. 3, no. December, pp. 87–93, 2019.

Kabir Ahmed, M., Maksha Wajiga, G., Vachaku Blamah, N., Modi, B., 2019. Stock market forecasting using ant colony optimization based algorithm. Am. J. Math. Comput. Model. 4 (3), 52.
Kaushik, S., et al., 2020. AI in Healthcare: time-series forecasting using statistical, neural, and ensemble architectures. Front. Big Data 3 (March).
Li J, Liu Y, Gong H, Huang X, "Stock price series forecasting using multi-scale modeling with boruta feature selection and adaptive denoising," Appl. Soft Comput., vol. 154, no. September 2023, p. 111365, 2024.
Lu, W., Li, J., Wang, J., Qin, L., 2021. A CNN-BiLSTM-AM method for stock price prediction. Neural Comput. Appl. 33 (10), 4741–4753.
Ma, Y., Mao, R., Lin, Q., Wu, P., Cambria, E., 2023. Multi-source aggregated classification for stock price movement prediction. Inf. Fusion 91 (April 2022), 515–528.
M. R. G., Panchal AT, "A hybrid strategy using mean reverting indictor Donchian channel and RSI," IOSR J. Econ. Financ. , vol. 16, no. 5, pp. 11–22, 2021.
Pabuccu H, Barbu A, "Feature Selection for Forecasting," pp. 1–21, 2023.
Pandya, J.B., Jaliya, U.K., 2021. Opinion and technical indicator based optimized deep learning for prediction of stock market. Indian J. Comput. Sci. Eng. 12 (6), 1860–1874.
Ratchagit, M., Xu, H., 2022. A two-delay combination model for stock price prediction. Mathematics 10 (19), pp.
Raudys, A., 2014. Optimal negative weight moving average for stock price series smoothing. IEEE/IAFE Conf. Comput. Intell. Financ. Eng. Proc. 1 (2), 239–246.
Rifai AD, "Comparison of implementation between EMA , WMA , SMA IN PREDICTING IHSG," vol. 2, no. 4, pp. 921–929, 2024.
Sarıkoç M, Celik M. PCA-ICA-LSTM: A Hybrid Deep Learning Model Based on Dimension Reduction Methods to Predict S&P 500 Index Price, no. 0123456789. Springer US, 2024.
Shohan, M.J.A., Faruque, M.O., Foo, S.Y., 2022. Forecasting of electric load using a hybrid LSTM-neural prophet model. Energies 15 (6), pp.
Tsilingeridis, O., Moustaka, V., Vakali, A., 2023. Design and development of a forecasting tool for the identification of new target markets by open time-series data and deep learning methods. Appl. Soft Comput. 132, 109843.
Vijh, M., Chandola, D., Tikkiwal, V.A., Kumar, A., 2020. Stock closing price prediction using machine learning techniques. Procedia Comput. Sci. 167 (2019), 599–606.
Walugembe, F., Stoica, T., 2022. Evaluating triple moving average strategy profitability under different market regimes. SSRN Electron. J.
Wang, S., 2023. A stock price prediction method based on BiLSTM and improved transformer. IEEE Access 11 (July), 104211–104223.
Wen, Y., Lin, P., Nie, X., 2020. "Research of stock price prediction based on PCA-LSTM model". IOP Conf. Ser Mater. Sci. Eng. 790 (1), pp.
Wu, J., Ren, P., Song, B., Zhang, R., Zhao, C., Zhang, X., 2023. Data glove-based gesture recognition using CNN-BiLSTM model with attention mechanism. PLoS One 18 (11 (November)), 1–22.
Yadav, K., Yadav, M., Saini, S., 2022. Stock values predictions using deep learning based hybrid models. CAAI Trans. Intell. Technol. 7 (1), 107–116.
"Yahoo Finance - Stock Market Live, Quotes, Business & Finance News." [Online]. Available: https://finance.yahoo.com/. [Accessed: 13-May-2024].
Yang, F., Chen, J., Liu, Y., 2023. Improved and optimized recurrent neural network based on PSO and its application in stock price prediction. Soft Comput. 27 (6), 3461–3476.
Zhang, H., 2018. The forecasting model of stock price based on PCA and BP neural network. J. Financ. Risk Manag. 07 (04), 369–385.
Z. Zhang, B. Chen, S. Zhu, and N. Langrené, "From attention to profit: quantitative trading strategy based on transformer," pp. 1–25, 2024.