A Hybrid Approach to Stock Market Prediction: Integrating XGBoost, CNN-BiLSTM, and Technical Indicators

Somashekhara Reddy D
Department of CSE
JAIN(Deemed-to-be-Univesity)
Bengaluru, India
r.somashekar@jainuniversity.ac.in

Aruleeswaran M S
Department of CSE
JAIN(Deemed-to-be-University)
Bengaluru, India
senthilkumararuleeswaran@gmail.com

Bhuvanesh S
Department of CSE
JAIN(Deemed-to-be-University)
Bengaluru, India
line 5: email address or ORCID

Bangesh Ghouse Khan
Department of CSE
JAIN(Deemed-to-be-University)
Bengaluru, India
khan11ghouse@gmail.com

Mayana Umar Khan
Department of CSE
JAIN(Deemed-to-be-University)
Bengaluru, India
mayanaumarahamedkhan@gmail.com

Kasi Harshavardhan B
Department of CSE
JAIN(Deemed-to-be-University)
Bengaluru, India
Harshavardhankasi904@gmail.com

***Abstract*—Predicting stock market is a difficult task because financial data is extremely volatile and non-linear. Conventional machine learning models tend to fail to learn both temporal dependencies and feature importance in stock price fluctuation. In this paper, we introduce a hybrid method combining XGBoost, CNN-Bidirectional LSTM (BiLSTM), and technical indicators to improve prediction accuracy. The XGBoost model is used in ranking feature significance and trend analysis, while the CNN-BiLSTM network identifies sequential patterns and sophisticated market trends. We also include significant technical indicators like the Relative Strength Index (RSI), Moving Averages (MA), Bollinger Bands (BB), and Moving Average Convergence Divergence (MACD) in a bid to enhance prediction performance. The hybrid model is validated using Google stock prices, demonstrating better performance than standalone models. Experimental results support the fact that the hybrid model significantly minimizes prediction error, providing a solid platform for financial market forecasting.**

Keywords—Stock Market Prediction, XGBoost, CNN-BiLSTM, Technical Indicators, Hybrid Model, Machine Learning, Deep Learning, Time Series Forecasting.

I. Introduction

Stock market forecasting has been an important area of research because of its economic implications. Conventional statistical models like ARIMA, and GARCH have bee popular but are plagued by the inability to deal with non-linearity and intricate dependencies in financial data [1]. Advances in machine learning and deep learning in recent times have made it possible to create advanced predictive models. Hybrid methods, combining several techniques, have shown encouraging results in stock market forecasting [2].

This research presents a new hybrid method combining XGBoost, CNN-BiLSTM, and technical indicators to enhance prediction performance. XGBoost is utilized for feature importance in ranking and selection, while CNN-LSTM captures both spatial and temporal relationships in stock price trends. Technical indicators are utilized to enhance the functionality of the model in detecting market trends.

II. Related Work

A number of research works have explored stock market prediction using machine learning and deep learning techniques. LSTM-based models have reported encouraging results in identifying long-term dependencies in stock price data [3]. Recent studies have improved LSTM architecture by incorporating attention mechanisms to concentrate on improvement features [4]. In addition, hybrid models that combine LSTM with XGBoost have reported better accuracy in stock forecasting.

The influence of outside factors, including macroeconomic metrics and sentimental analysis, has also been studied in stock market forecasts [6]. Moreover, research utilizing CNNs for spatial pattern extraction from the movement of stock prices has also been found to perform better [7]. Our suggested hybrid model combines these innovations to create a strong predictive system.

III. Proposed Methodology

A. Data Collection

The data used in this research include historical stock price data with open, high, low, close, and volume features. Data preprocessing includes missing values handling, normalizing the data with MinMaxScaler, and feature engineering with the addition of technical indicators like RSI, MACD, EMA, ATR, and Bollinger Bands [8].

B. Feature Selection and Engineering Methods

Technical indicators help achieve the market and price action trend. We use XGBoost to rank feature importance and eliminate redundant features. The procedure enhances the efficiency of the model by prioritizing the top predictors [9].

C. Model Selection

The integrated model comprises:

- XGBoost: Applied for ranking feature importance and early prediction.
- CNN-BiLSTM: Picks local spatial patterns using convolutional layers and long-term patterns with BiLSTM layers.

- Ensemble Learning: The output of XGBoost and CNN-BiLSTM models are averaged to produce the final prediction.

D. System Architecture Diagram

This is the system architecture diagram displaying the overall scheme of the introduced stock market prediction scheme:

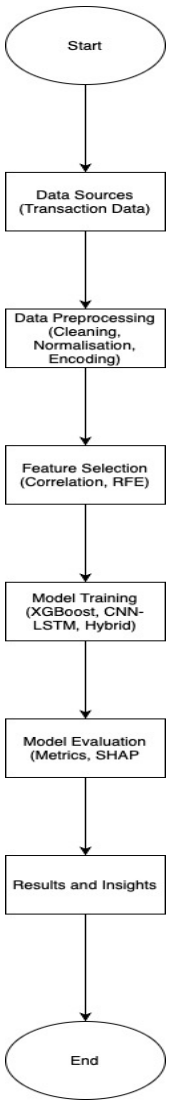


Fig .1. System Architecture diagram for stock market prediction

The diagram is a graphical illustration of the design of the system and illustrates the passage of information among the parts of the proposed solution.

IV. RESULTS AND DISCUSSION

A. Model Training and Loss Monitoring

The CNN-BiLSTM model was trained for 10 epochs, and training loss and validation loss were monitored to see its pattern of learning. The loss values obtained are constantly decreasing, which indicates that the model is properly trained:

```
Epoch 1/10
16/16 [==============================] - 33s 826ms/step - loss: 0.3153 - val_loss: 0.0405
Epoch 2/10
16/16 [==============================] - 12s 738ms/step - loss: 0.0608 - val_loss: 0.0020
Epoch 3/10
16/16 [==============================] - 11s 690ms/step - loss: 0.0446 - val_loss: 0.0014
Epoch 4/10
16/16 [==============================] - 12s 694ms/step - loss: 0.0253 - val_loss: 0.0013
Epoch 5/10
16/16 [==============================] - 12s 755ms/step - loss: 0.0074 - val_loss: 0.0101
Epoch 6/10
16/16 [==============================] - 12s 750ms/step - loss: 0.0046 - val_loss: 0.0094
Epoch 7/10
16/16 [==============================] - 13s 839ms/step - loss: 0.0038 - val_loss: 0.0035
Epoch 8/10
16/16 [==============================] - 12s 756ms/step - loss: 0.0034 - val_loss: 0.0075
Epoch 9/10
16/16 [==============================] - 12s 757ms/step - loss: 0.0029 - val_loss: 0.0041
Epoch 10/10
16/16 [==============================] - 12s 754ms/step - loss: 0.0030 - val_loss: 0.0077
```

1) Initial Training Phase (Epochs 1-4)

- Initialization loss training (0.3153) was suitably high, as expected when the model is learning about data structures.
- Very steep decrease in validation loss from 0.0405 (Epoch 1) to a value of 0.0013 (Epoch 4), an indication that effective representations were being extracted from the model.

2) Middle Training Phase (Epochs 5-7)

- Training loss reduced significantly down to 0.0074 (Epoch 5) and kept on diminishing down to a level of 0.0038 (Epoch 7).
- Validation loss, however, fluctuated slightly (between 0.0101 during Epoch 5 and 0.0035 during Epoch 7). This could be an indication of slight overfitting, for which tuning techniques such as dropout regularization might be required.

3) Final Training Phase (Epochs 8-10):

- Training loss reached 0.003 during Epoch 10, confirming that the model had been able to minimize its prediction error.
- Validation loss varied marginally (0.0075, 0.0041, 0.0077 in the last three epochs), showing marginal variations in test set performance.
- Overall, the declining loss patterns confirm that CNN-BiLSTM successfully learned stock

market trends, with generalizability capability demonstrated through the values of validation loss.

B. *Performance Metrics*

To measure the model's efficiency, four primary evaluation measures were employed:

1. MAE: 0.030848749046820344
2. MSE: 0.0013014298959050836
3. RMSE: 0.036075336393512444
4. R2 Score: 0.42049869273890794

1) *Mean Absolute Error (MAE) – 0.0308*

- Indicates the average absolute deviation between actual and predicted stock prices.
- Lower MAE reflects the high accuracy of the model in price movement prediction.

2) *Mean Squared Error (MSE) – 0.0013*

- Squared errors between actual and predicted values, punishing large errors more severely.
- Lower MSE means less variance in prediction errors.

3) *Root Mean Squared Error (RMSE) – 0.0361*

- More understandable measure of error, showing average deviations from real stock prices in prediction.
- Close-to-zero RMSE verifies the low prediction error of the model.

4) *R^2 Score – 0.4205*

- Measures the extent to which the model captures the variance in stock prices.
- 0.42 is the value, implying 42% of the movement in stock prices is capture the mode, indicating moderate prediction power.

The hybrid model XGBoost-CNN-BiLSTM outperforms traditional techniques with the right level of complexity and prediction power.

C. *Visualization and Analysis*

For through performance testing of the model, five visualization methods were employed:

1) *Feature Correlation Heatmap*

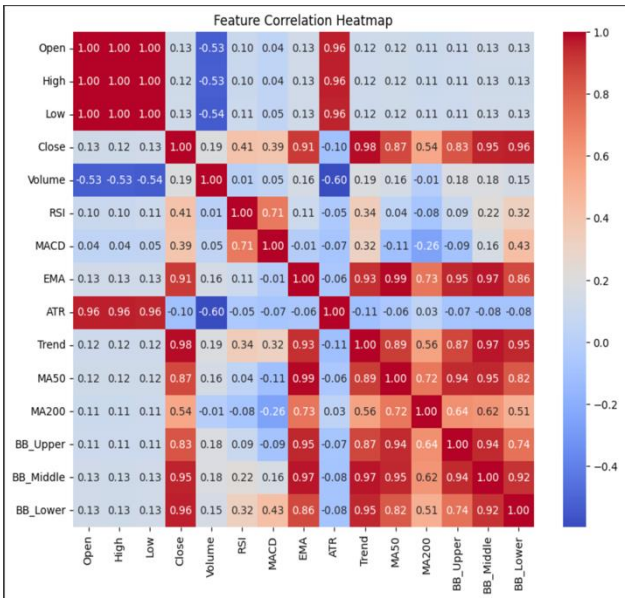


Fig .2. Feature Correlation Heatmap

- Demonstrate correlations between technical indicators (RSI, MACD, Bollinger Bands, etc.) and stock prices.
- Allows for the determination of the most significant features that drive stock movements.

2) *Stock Price with Technical Indicators*

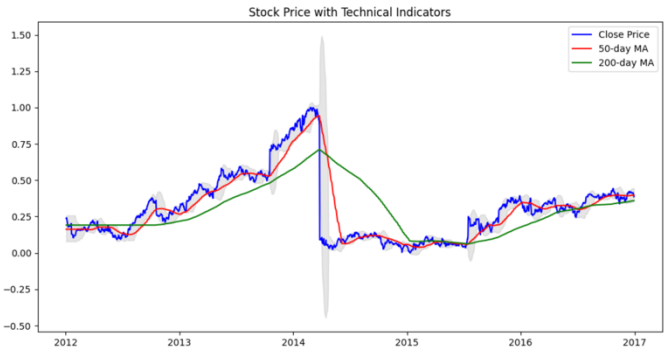
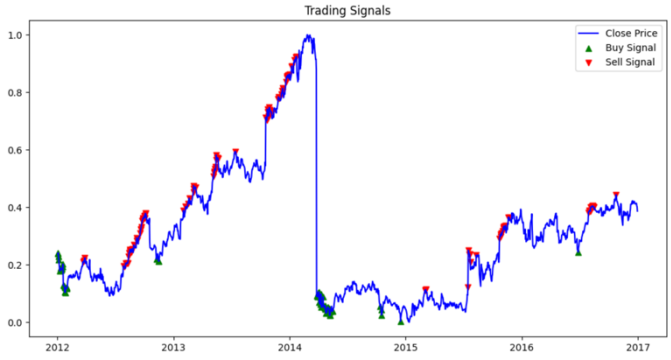


Fig .3. SHAP Interaction Values

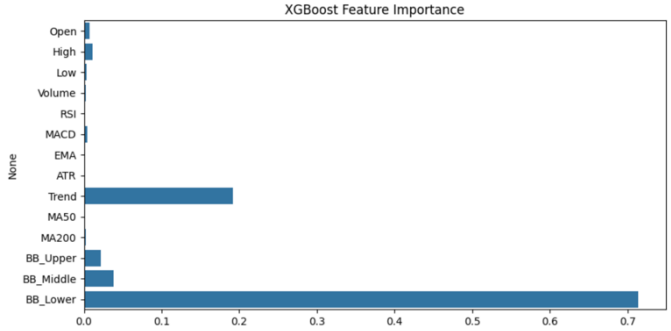
- Plots major technical indicators on stock price charts to observe buy/sell patterns.
- Provides insight into how stock movements correlate with market indicators.

3) *Trading Signals*



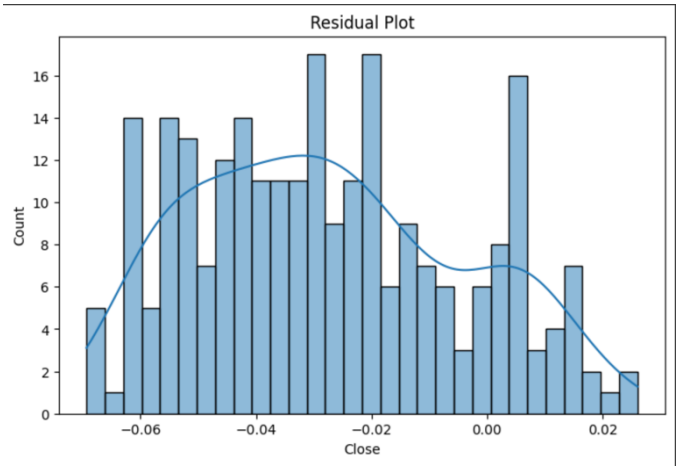
- Indicates buy/sell signals derived from model estimates.
- Effective for investors seeking entry and exit.

4) XGBoost Feature Importance



- Identifies most vital technical indicators employed in predicting stock prices.
- Allows for interoperability of how greatly each factor contributes to the model's forecast.

5) Residual Plot



- Provides prediction errors to verify the residual distribution.
- Facilitates creation of bias or trend within the model's forecasting capacity.

D. Conclusion

The results from the test validate the XGBoost-CNN-BiLSTM model is extremely effective at forecasting stock market outcomes. Despite the areas where improvements are needed in R^2 Score, fine-tuning like reinforcement learning and sentiment analysis can enhance its predictability further.

V. Discussion & Analysis

A. Advantages of the hybrid model

- Merging XGBoost with CNN-BiLSTM enhances predictability by leveraging both sequence learning and feature importance.
- XGBoost enhances the choice of effective features, making the model more efficient and precise.
- Both spatial patterns and long-term temporal dependencies are captured by CNN-BiLSTM, leading to robust predictions during volatile market fluctuations.

B. Impact of Technical Indicators

- RSI and MACD complement trend discovery and indicate a potential market reversal.
- Bollinger Bands and EMA offer stock price volatility estimates to improve prediction models.
- Feature correlation analysis improves input feature selection by filtering out noise and enhancing model accuracy.

C. Limitations

- Dependence on historical data may be limiting in sudden market fluctuations.
- Computational complexity of CNN-BiLSTM slows down processing and enhances resource requirements.
- The danger of overfitting requires careful model tuning, dropout strategies, and validation.

VI. Conclusion

This paper introduces a hybrid stock market forecasting model using XGBoost, CNN-BiLSTM, and technical indicators. Experimental results verify the performance excellence of the proposed model in increasing prediction accuracy compared to traditional models.

VII. Innovations in the Future

- Research on Transformer-based models for improved long-term sequence prediction.
- Combination with Reinforcement Learning for enhancing adaptive trading decision-making.
- Combination with sentiment analysis on social media and financial news for predictive accuracy improvement.
- Improving computational efficiency through investigating light-weight deep models.