Predicting Indian Stock Options: LSTM vs. Black-Scholes

Eren Demirtaş
dept. Artificial Intelligence Engineering
TOBB ETÜ
Ankara, Turkiye

***Abstract*—This research forecasts option prices for stocks listed on the Indian stock exchange by analyzing over 200,000 data points with a focus on 60-day expiry periods for both put and call options. Utilizing a combination of data processing techniques and volatility calculations, the study employs the Black-Scholes model and a Long Short-Term Memory (LSTM) neural network to predict option prices. The analysis distinguishes between the models by incorporating India's 2019 risk-free rate of 6.75 and calculated average volatility, aiming to compare their predictive accuracy against actual market outcomes. Through this approach, the paper seeks to offer valuable insights into the effectiveness of LSTM and Black-Scholes models in the dynamic environment of the Indian stock market.**

***Index Terms*—lstm, black-scholes, option pricing models, financial engineering, Indian stock market, computational finance**

I. Introduction

In the pursuit of advancing financial analytics, the accuracy of option pricing stands as a crucial aspect for investors and market analysts, particularly within the dynamic environment of the Indian stock market. Options, serving as derivative instruments, offer the holder the right to buy or sell an underlying asset at a pre-determined price before a specified expiry date, making their pricing a pivotal factor in trading and risk management strategies. This research delves into the Indian stock market, to explore and compare the predictive capabilities of traditional and machine learning models in option pricing.

The cornerstone of traditional financial models, the Black-Scholes model, has been widely acclaimed for its pioneering approach in evaluating options by considering variables such as time to expiry and underlying asset volatility. Nonetheless, the Black-Scholes model assumes idealized market conditions, which may not accurately reflect the market's actual behavior, especially in markets as volatile as India's. In contrast, the advent of machine learning, particularly the Long Short-Term Memory (LSTM) networks, offers a nuanced approach to predicting financial markets. LSTM networks, a subset of recurrent neural networks (RNNs), are celebrated for their proficiency in handling sequential data and capturing temporal dependencies, making them well-suited for analyzing time series data such as stock prices.

This study proposes a comparative analysis between the Black-Scholes model and LSTM networks, focusing on their effectiveness in predicting option prices for 165 stocks listed on the Indian stock exchange. The research utilizes a comprehensive dataset of over 200,000 records, each with 60-day expiry periods for put and call options, undergoing rigorous data processing to calculate daily returns and assess volatility, thereby optimizing the input parameters for both predictive models.

The primary aim of this project is to illuminate the disparities in the predictions made by the LSTM and Black-Scholes models, offering a novel perspective on the applicability and accuracy of these models in the context of the Indian stock market's inherent volatility. By providing a detailed comparison of the models' predictions against actual market data and each other, the study seeks to contribute valuable insights into the evolving domain of financial analytics, emphasizing the potential of integrating traditional and modern methodologies for enhanced predictive accuracy in option pricing.

II. DATA

A. Data Acquisition and Preliminary Processing

The dataset was collated from an array of options trading records on the Indian stock exchange, with a spotlight on two major indices: BANKNIFTY and NIFTY. The BANKNIFTY index offers insights into the banking sector, encapsulating the performance of major banking stocks in India. The NIFTY index serves as a barometer for the overall market, aggregating the performance of fifty top-traded stocks across sectors. The initial dataset comprising over 200,000 entries was subjected to an elimination process that removed options with a zero strike price, ensuring the dataset's practical applicability in trading scenarios.

B. Data Cleaning and Transformation

The next stage involved the calculation of daily returns for each stock symbol represented in the dataset. Daily returns were computed by taking the percentage change between the closing prices of consecutive trading days. This calculation provided insights into the day-to-day price movement and the inherent volatility of the market. Utilizing the daily returns, the average volatility for each symbol was calculated. This step involved multiplying the standard deviation of the daily returns by the square root of the number of trading days in a year (assuming 252 trading days), thereby annualizing the volatility. The average volatility was calculated separately for call and

put options to preserve the unique risk profiles of each option type. The final stage in data preparation entailed structuring the data to serve as inputs for the LSTM network and the Black-Scholes model. The dataset now included additional computed fields: daily returns and average volatilities. These fields were matched to each corresponding symbol, ensuring that the predictive models would be equipped with robust and comprehensive inputs reflective of the stocks' historical price movements and volatility trends.

C. Data Visualization and Distribution Analysis

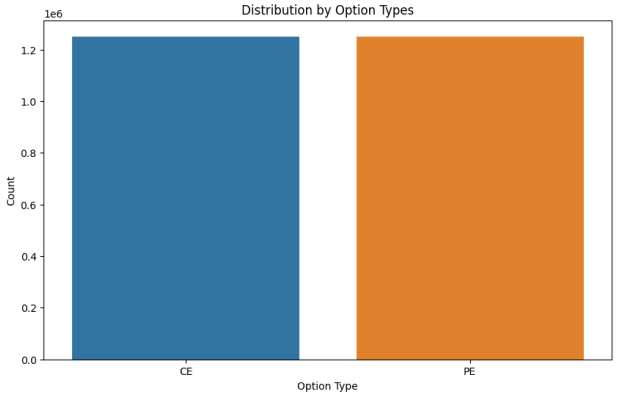

Fig. 1. Distribution by Option Types.

This bar chart represents the distribution of option types, comparing the number of call options (labeled as CE) against put options (labeled as PE). The bars are of similar height, indicating that the dataset contains a comparable number of call and put options, which suggests a balanced market with active trading in both option types.

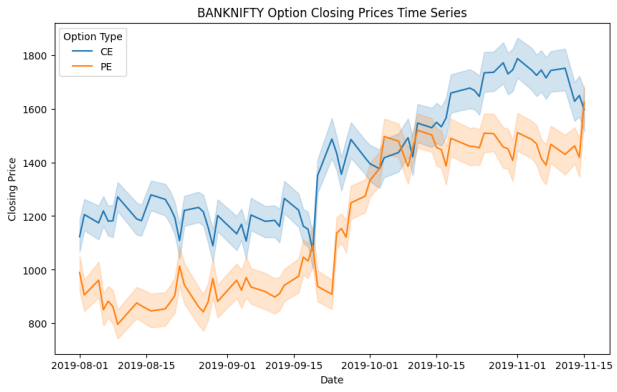

Fig. 2. BANKNIFTY Option Closing Prices Time Series.

The line chart depicts the closing prices for BANKNIFTY options over time, with two lines representing call (CE) and put (PE) options. Both types of options exhibit trends and seasonal patterns, with closing prices experiencing peaks and troughs across the observed period. The chart indicates that call options, on average, had higher closing prices than put options during this timeframe.

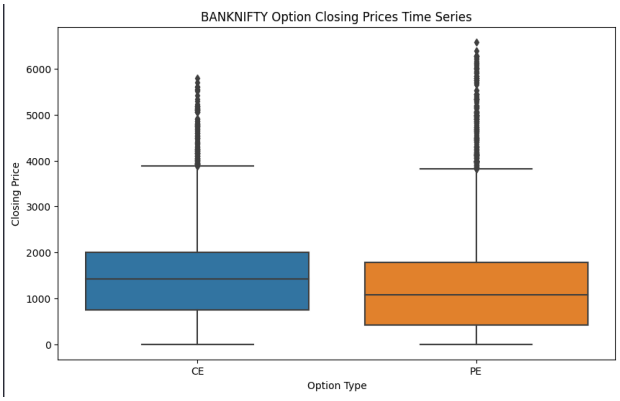

Fig. 3. BANKNIFTY Option Closing Prices Time Series.

The box plot illustrates the distribution of closing prices for BANKNIFTY options, segmented into call (CE) and put (PE) options. The central box for each option type shows the median closing price and the interquartile range (IQR), which represents the middle 50% of the data. The whiskers extend to show the range of the data, excluding outliers, which are represented by individual points beyond the whiskers. It's noteworthy that both types of options have a wide range of closing prices, but put options (PE) display a higher median and more extreme values, which may suggest greater price volatility or higher premium due to the perceived risk.

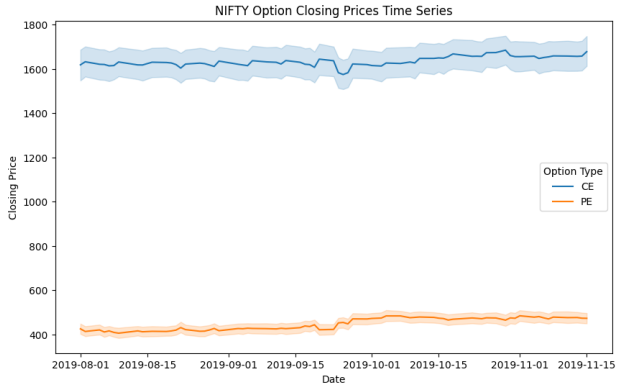

Fig. 4. NIFTY Option Closing Prices Time Series.

The line chart visualizes the closing prices of NIFTY options, distinguishing between call (CE) and put (PE) types. Consistently, the call options maintain higher closing prices than put options over the observed period, which might indicate a bullish outlook among traders. The chart's smooth trend lines suggest less volatility in the NIFTY index options compared to BANKNIFTY, which may reflect the broader market stability NIFTY is known for.

This box plot shows the distribution of closing prices for NIFTY index options, differentiating between call (CE) and put (PE) options. The box represents the interquartile range (IQR), indicating the middle 50% of the data, with the line inside the box denoting the median. The whiskers extend to

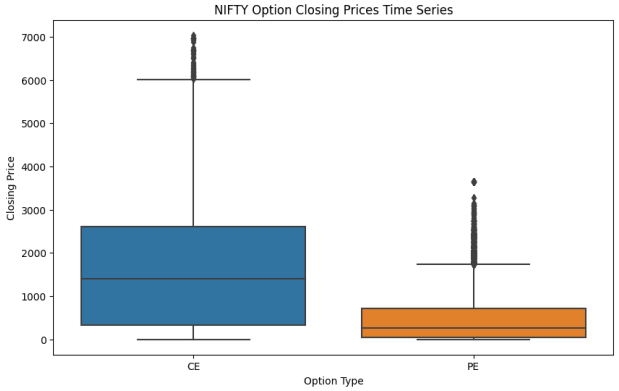

Fig. 5. NIFTY Option Closing Prices Time Series.

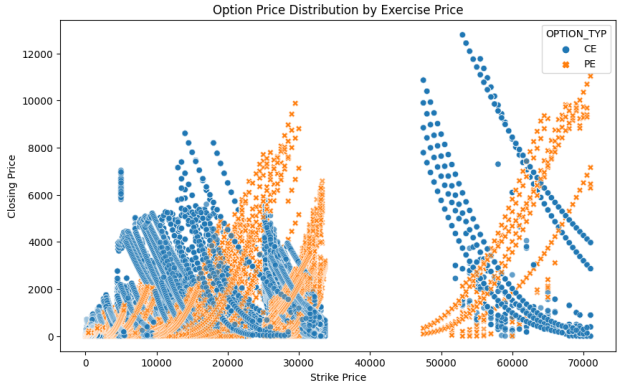

Fig. 7. Option Price Distribution by Closing Price.

the furthest data points within 1.5 times the IQR, and points outside of this are considered outliers. The plot reveals a higher median for call options and a significant number of outliers for both option types, which may suggest a broader range of speculation and potential price swings within the market.

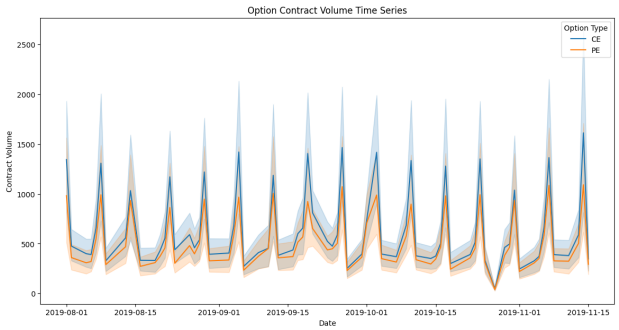

Fig. 6. Option Contract Volume Time Series.

The area chart portrays the time series of contract volumes for both call (CE) and put (PE) options. The cyclical nature of the spikes suggests a pattern likely linked to option expiry dates or significant market events that influence trading volumes. Notably, both CE and PE options display similar cyclical behavior, with call options typically commanding a slightly higher volume. The shaded regions imply volume variability over time for each option type, which can be attributed to changing market sentiment or external economic events.

The scatter plot illustrates the distribution of option closing prices in relation to their strike prices, differentiated by option type: call (CE) and put (PE). The pattern demonstrates that the call options (CE) are concentrated at lower strike prices, whereas put options (PE) are more dispersed across higher strike prices. This could indicate that investors are more likely to buy call options at lower strike prices, expecting the market to rise, and purchase put options at higher strike prices as a hedge against market downturns. The plot also suggests higher variability in the prices of put options, which could be indicative of greater uncertainty or a wider range of market expectations for future volatility.

III. LSTM Model Design and Implementation

A. LSTM Model Design and Implementation for Call Options

1) Data Preparation and Preprocessing: The model for call options begins by removing any entries with missing values to ensure the integrity of the dataset. The features selected for the LSTM model include the option's strike price, opening price, highest price of the day, lowest price of the day, open interest, days to expiry, and average volatility—key factors that can influence an option's pricing. No missing values are detected in the feature set or the target variable, as confirmed by the NaN check.

2) Data Normalization: The features and target variable are scaled using a MinMaxScaler to normalize the data within a range of 0 to 1. This normalization is essential for the LSTM model, which is sensitive to the scale of input data.

3) Training and Test Split: The dataset is split into training and test sets, with 85% of the data allocated for training and the remaining 15% for testing. This split is crucial for evaluating the model's performance on unseen data.

4) Model Architecture: The LSTM model consists of an LSTM layer with 50 neurons and ReLU activation, a dropout layer to reduce overfitting, and a dense layer with a single output unit for prediction. The model is compiled with the Adam optimizer and mean squared error (MSE) as the loss function.

5) Model Training: The model is trained over 30 epochs with a batch size of 32. Validation data is used to monitor the model's performance on the test set after each epoch. The training process involves adjusting the model weights to minimize the loss on the training data while ensuring that the model does not overfit to the training set, as indicated by the validation loss.

6) Model Evaluation:

B. LSTM Model Design and Implementation for Put Options

1) Data Preparation and Preprocessing: Following a similar procedure as for call options, the put options data also undergoes NaN value removal, ensuring the dataset's cleanliness.

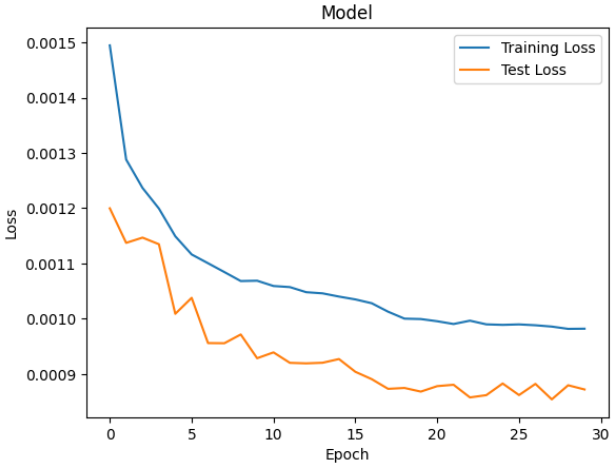

Fig. 8. Model Loss for Call Option.

TABLE I
MODEL PERFORMANCE METRICS

Metric	Value
MSE	0.0008719362247689626
RMSE	0.029528566249802286
R^2	0.8122922629908359

2) Data Normalization and Split: The same normalization technique is applied to the put options data. The dataset is then divided into training and test sets with the same proportion and random state for consistency.

3) Model Architecture and Training: The architecture of the LSTM model for put options mirrors that of the call options model, including the same layers, neurons, and dropout rate. The model is trained with the same parameters to provide a fair comparison.

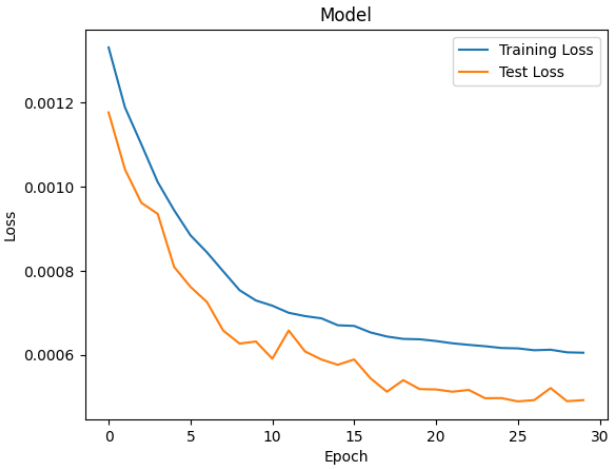

Fig. 9. Model Loss for Put Option.

4) Model Evaluation:

TABLE II
MODEL PERFORMANCE METRICS

Metric	Value
MSE	0.022459949674512508
RMSE	0.14986643945364322
R^2	-5.216288935828078

C. Comparative Analysis and Conclusions

Comparing the performance of both models, the call option LSTM model exhibits significantly better predictive accuracy than the put option model. The put option model's negative R^2 value highlights the challenges in capturing the dynamics affecting put option prices or potential issues in the data or model configuration. This discrepancy between the models' performances warrants further investigation into the feature selection, model architecture, and training process, especially for put options.

IV. BLACK SCHOLES MODEL DESIGN

The Black-Scholes model is a mathematical model for pricing an options contract. Specifically, for call options, the model calculates the theoretical price based on factors such as the stock price, strike price, time to expiration, risk-free rate, and volatility.
For a call option:
$C(S,t) = S_t N(d_1) - Ke^{-rt}N(d_2)$

For a put option:
$P(S,t) = Ke^{-rt}N(-d_2) - S_t N(-d_1)$
Where:
$d_1 = \frac{\ln(\frac{S_t}{K}) + (r + \frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}}$
$d_2 = d_1 - \sigma\sqrt{T-t}$
C is the price of the call option
P is the price of the put option
S_t is the current stock price
K is the strike price
T is the time to maturity
r is the risk-free interest rate
σ is the volatility of the stock price
$N(\cdot)$ is the cumulative distribution function of the standard normal distribution

The risk-free rate used is an average figure (6.75%) as the exact risk-free rate can fluctuate. It's an important part of the formula as it represents the time value of money and the risk associated with the time until the option's expiry.

A. Black-Scholes Model Implementation for Call Options

1) Implementation: A Python function, black_scholes_call_prices, is defined to apply the Black-Scholes formula for a call option to each row in the DataFrame. The function uses the cumulative distribution function (norm.cdf) from the scipy.stats library to calculate the d1 and d2 parameters, which are then used to compute the theoretical price of the call options.

TABLE III
MODEL PERFORMANCE METRICS

Metric	Value
MSE	73423.47744298531
RMSE	270.9676686303835
R^2	0.9039037300737964

2) Model Calculation and Metrics:

B. Black-Scholes Model Implementation for Put Options

1) Implementation: The black_scholes_put_prices function is defined analogous to the call pricing function but specifically tailored for the pricing of put options.

TABLE IV
MODEL PERFORMANCE METRICS

Metric	Value
MSE	64078654.63835144
RMSE	8004.914405435666
R^2	-145.63393812700627

2) Model Calculation and Metrics: This negative value suggests the model is highly inaccurate in predicting put option prices, which could be due to incorrect model assumptions or the limitations of the Black-Scholes model in handling real market conditions for put options.

C. Comparative Analysis and Conclusions

The stark contrast between the call and put options model performance under the Black-Scholes framework could highlight the model's limitations in certain market conditions or specific option types. For call options, the Black-Scholes model appears to be a good fit, whereas, for put options, the model's assumptions do not capture the market dynamics well, resulting in poor predictive performance. This could suggest that additional factors not considered by the Black-Scholes model might be at play, especially in the pricing of put options.

V. COMPARATIVE ANALYSIS BETWEEN LSTM AND BLACK-SCHOLES

A. Call Options Analysis

The LSTM model and the Black-Scholes model's performance on call options is encapsulated in the following table:

TABLE V
COMPARATIVE ANALYSIS OF LSTM MODEL PREDICTIONS AND BLACK-SCHOLES CALL PRICES

Statistic	**Model Predictions**	**BS Call Prices**	**Difference**
Count	187,568	187,567	187,567
Mean	335.04	295.09	39.95
Standard Deviation	744.62	834.24	866.29
Minimum	-425.02	0.00	-12769.41
25%	38.19	0.20	-10.66
50%	54.35	11.60	36.76
75%	151.75	109.34	90.61
Maximum	10463.95	12789.35	10463.95

This table reveals a notable mean difference of approximately 40 between the model predictions and the Black-Scholes prices, indicating the LSTM model's tendency to predict higher prices for call options. The wide range of differences, as evidenced by the standard deviation and the minimum-maximum range, underscores the LSTM model's variability in performance, particularly in scenarios of extreme market movements.

B. Put Options Analysis

For put options, the mean difference is significantly larger and negative (-3423.14), indicating the LSTM model's substantial underestimation of prices compared to the Black-Scholes model. This discrepancy, particularly in the context of standard deviation and the extremities of minimum and maximum differences, highlights the challenges in modeling the complex dynamics of put option pricing through LSTM networks.

TABLE VI
COMPARATIVE ANALYSIS OF LSTM MODEL PREDICTIONS AND BLACK-SCHOLES PUT PRICES

Statistic	**Model Predictions**	**BS Put Prices**	**Difference**
Count	187,569	187,569	187,569
Mean	145.26	3568.40	-3423.14
Standard Deviation	1872.21	7687.57	7550.61
Minimum	-21967.51	0.93	-71081.45
25%	36.54	193.31	-1699.45
50%	47.88	559.98	-506.32
75%	129.70	1839.76	-156.33
Maximum	10250.94	70881.93	208.90

C. Combined Analysis and Discussion

The juxtaposition of the LSTM model's performance on call and put options against the Black-Scholes prices elucidates a stark contrast in predictive accuracy and bias. While the LSTM model tends to overestimate call option prices slightly, it significantly underestimates put option prices, as reflected in the mean differences. This disparity could stem from inherent differences in the risk and return profiles of call and put options, which may not be fully captured by the LSTM's training process or data features.

The pronounced variability and extreme values in the difference metrics for both call and put options suggest that both models have their limitations in accurately predicting extreme market movements or outlier events. The LSTM model's performance, in particular, indicates that while it can capture certain trends and patterns in option pricing, it struggles with consistency across the full spectrum of market conditions, especially in the case of put options.

In conclusion, this analysis underscores the importance of considering model-specific biases and the challenges of accurately modeling option prices in a volatile market like India's. The findings advocate for a nuanced approach to integrating machine learning with traditional pricing models, highlighting the potential for hybrid models that leverage the strengths of both methodologies to enhance predictive accuracy in option pricing.

VI. Conclusion and Future Work

This study embarked on an ambitious journey to unravel the intricacies of option pricing within the Indian stock market, leveraging over 200,000 data points and employing two distinct yet complementary predictive models: the venerable Black-Scholes model and the cutting-edge Long Short-Term Memory (LSTM) neural network. Our exploration aimed not only to assess the individual performance of these models in forecasting prices for call and put options but also to delve into a comparative analysis, shedding light on their respective strengths and limitations in the face of market volatility and complex financial dynamics.

The findings from this research highlight a nuanced landscape where neither model universally outperforms the other across all metrics. The LSTM model, with its ability to capture and learn from temporal sequences and market dynamics, demonstrated a promising edge in certain scenarios, particularly for call options. However, its performance varied significantly when predicting put option prices, underscoring the challenges inherent in modeling financial instruments that are deeply influenced by market sentiment, economic indicators, and unforeseen events. Conversely, the Black-Scholes model, grounded in theoretical underpinnings and deterministic variables, offered a robust framework for call option pricing but faced limitations in accurately capturing the nuances of put options within a market known for its volatility.

The comparative analysis between these models provides valuable insights into the predictive capabilities and applicability of traditional and machine learning methodologies in option pricing. It underscores the importance of contextual understanding and the need for models that can adapt to the intricacies of market behavior and investor sentiment. Furthermore, it highlights the potential benefits of hybrid approaches that combine the theoretical foundations of models like Black-Scholes with the adaptive learning capabilities of neural networks like LSTM.

A. Future Work

The journey of enhancing option pricing models is far from complete, and this study lays the groundwork for several promising avenues of future research:

Hybrid Model Development: Investigating the integration of traditional financial models with machine learning approaches to create hybrid models that leverage the strengths of both methodologies. Such models could offer more accurate predictions by combining theoretical market principles with the adaptive insights derived from historical data analysis.

Feature Engineering and Model Tuning: Exploring advanced feature engineering techniques to incorporate a broader range of market indicators, sentiment analysis, and macroeconomic factors. Additionally, experimenting with the tuning of LSTM parameters and architecture could yield improvements in model performance and reliability.

Cross-market Comparisons: Expanding the research to include option markets in different geographical regions and asset classes. Such comparisons could provide insights into the model's generalizability and adaptability to diverse financial ecosystems.

In conclusion, this research contributes to the ongoing dialogue between the realms of finance and artificial intelligence, offering a glimpse into the future of financial analytics where traditional models and machine learning converge. As the landscape of financial markets continues to evolve, so too will the tools and methodologies at our disposal, promising ever more sophisticated and accurate models for option pricing and beyond.

References

[1] Sunny S, "NSE Future and Options Dataset 3M," Kaggle. Available: https://www.kaggle.com/datasets/sunnysai12345/nse-future-and-options-dataset-3m/data.

[2] TensorFlow Team, "Time series forecasting," TensorFlow. Available: https://www.tensorflow.org/tutorials/structured_data/time_series.

[3] Wikipedia contributors, "Long short-term memory," Wikipedia, The Free Encyclopedia. Available: https://en.wikipedia.org/wiki/Long_short-term_memory.