Short Term Stock Prediction Under Low Data Constraints Using Hybrid Deep Learning

Aadi Srivastava
Computational Modeling and Data Analytics
Virginia Tech University
Blacksburg, VA 24060
aadisrivastava@vt.edu

Caleb Appiagyei
Computer Science - Data-Centric Computing
Virginia Tech University
Blacksburg, VA 24060
caleba04@vt.edu

Justin Morris
Computer Engineering - Software Systems
Virginia Tech University
Blacksburg, VA 24060
jgmorris@vt.edu

Andrew Kinzie
Computer Engineering - General
Virginia Tech University
Blacksburg, VA 24060
akinzie1@vt.edu

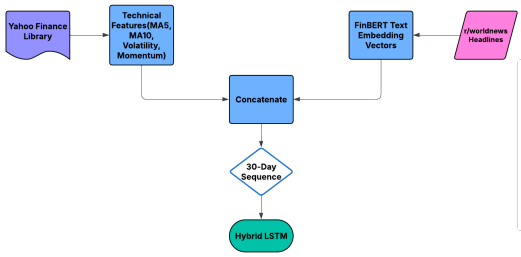
Abstract

Short term stock prediction remains a challenging problem for machine learning, especially under low data and constraints faced by everyday investors. While hybrid models that combine numerical signals with textual sentiment have shown promise, most evidence comes from high frequency price features or large scale news data, leaving open the question of whether such models retain value in realistic, resource limited environments. In this study, we systematically evaluate the performance boundaries of hybrid deep learning models under low resolution daily data. Using Yahoo Finance price indicators and aggregated Reddit world news(r/worldnews) headlines, we construct a multimodal dataset and compare classical baselines, Logistic Regression and Random Forest, LSTM, and a Hybrid LSTM integrating FinBERT embeddings. Our analysis shows that baseline models and price only sequence models perform nearly the same, whereas the Hybrid LSTM achieves the highest accuracy, F1 score, and AUC, demonstrating that combining news text embeddings with daily price trends yields measurable predictive gains even without high frequency signals. These results highlight that hybrid modeling remains effective within realistic public data constraints and may offer practical relevance for everyday users. Individual group contributions: Aadi developed the LSTM and Hybrid LSTM architectures and conducted deep learning experimentation. Caleb established the project's initial data environment, gathering and preparing raw stock data and news datasets. He helped train the baseline models, which would later serve as a benchmark for our hybrid architecture. Justin developed the full data processing and modeling pipeline that were used throughout the project. He was able to implement feature extraction workflow for price indicators and dataset merging between market data and aggregated news text. Andrew implemented the full evaluation framework, generating performance metrics, comparison tables, and visualizations used to analyze all models. He also produced the Hybrid LSTM results, created the model comparison figures, and led the development of the presentation materials and final report narrative, including the Discussion and Conclusion sections.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

Stock prediction remains an active challenge in machine learning, largely due to the noisy, fluctuating, and highly nonlinear structure of financial time series. Classical models such as, Logistic Regression and Random Forests, have long relied on engineered technical indicators to capture short term dynamics, yet these approaches struggle to learn temporal dependencies or incorporate external market sentiment. Recent advances in deep learning, particularly sequence models like LSTMs and transformer based sentiment encoders such as FinBERT, have started interest in hybrid models that combine numerical and textual data. These hybrid approaches have shown promising results in research settings using high frequency indicators, rich news data, or institution grade sentiment streams. Despite their success, most existing studies leave a critical question unexamined. Do hybrid models truly maintain their predictive advantages when constrained to the low data conditions that everyday investors face? Most prior studies do not clearly separate whether the improved performance comes from the hybrid model itself or simply from having large, highly detailed datasets. For example, many transformer augmented prediction systems operate over minute level price ticks or thousands of headlines per day. These conditions are far removed from how a retail trader, student, or small research team interacts with financial data. Similarly, while several works demonstrate improvements from sentiment embeddings, they typically rely on professionally curated financial news streams, not publicly aggregated and noisy sources such as Reddit headlines. This uncertainty hinders assessing the true practical value of hybrid modeling. If hybrid systems require massive datasets or high frequency signals, then their usefulness for small scale users is limited. Conversely, if hybrid architectures maintain performance benefits even when restricted to a single daily price bar and a small collection of publicly sourced headlines, then their utility extends far beyond the institutional domain.



	Model	Accuracy	F1-Score	AUC
0	Logistic Regression (Indicators)	0.511111	0.500000	0.510989
1	Random Forest (Indicators)	0.466667	0.601476	0.479853
2	LSTM (Price Sequence)	0.463659	0.610909	0.472776
3	Hybrid LSTM (Price + News)	0.548872	0.628099	0.558575

Figure 1: **Left:** Overview of the multimodal dataset used in our study, combining daily technical indicators from Yahoo Finance with aggregated Reddit WorldNews headlines encoded via FinBERT. The resulting numerical and textual features are concatenated and transformed into 30-day temporal windows for sequence modeling. **Right:** Performance comparison across baseline models, Logistic Regression and Random Forest, LSTM, and our Hybrid LSTM.

In this paper, we investigate whether hybrid deep learning models can extract meaningful predictive signal using only daily price indicators and aggregated Reddit world headlines (Fig. 1). Our exploration compares two classical indicator based baselines, Logistic Regression and Random Forest, a price only LSTM sequence model, and a Hybrid LSTM that integrates 30-day historical windows with FinBERT embeddings derived from daily news text. Consistent with our motivation, we intentionally adopt a low data regime including, a single closing price bar per day and the top 25 user ranked headlines sourced from a public Kaggle dataset. This setting allows us to isolate whether hybrid architectures provide intrinsic modeling advantages or merely capitalize on high resolution data availability. Our findings reveal that while baseline models and price only LSTMs perform similarly, hybrid models consistently outperform them, achieving improved accuracy, F1 score, and AUC even under strict data limitations. This suggests that the hybridization of numerical and textual signals contains intrinsic value, not just in data rich contexts, but also in low resolution environments that reflect real world constraints. Taken together, our results provide a deeper understanding of the conditions under which hybrid financial models offer practical benefit, and they highlight the feasibility of multimodal learning for everyday market participants.

2 Literature Review and Background

Classical machine learning approaches such as logistic regression and random forests have long been used for short term stock prediction, largely relying on manually engineered technical indicators rather than temporal or textual structure. However, recent work suggests that these traditional models face structural performance ceilings when applied to noisy, nonlinear financial time series. The Stanford study by Zheng and Jin [1] highlights this limitation directly. Although their model incorporated both daily technical indicators and sentiment features, their LSTM failed to improve performance, leading the authors to explicitly recommend deeper exploration of neural architectures capable of modeling sequential dependencies and news information more effectively. A second influential source for our study comes from Florida International University, which used AAPL stock data together with over 6,000 labeled tweets to build sentiment aware prediction models [2]. Their analysis showed that classical baselines underperformed relative to deep models, and they concluded that hybrid architectures that combine numerical and textual signals may represent the next meaningful advancement in financial prediction. This observation strongly shaped our motivation, particularly because their results imply that sentiment contributes measurable information even when the underlying price features are relatively simple. Building on these foundations, the work from IIIT expanded the scope of hybrid modeling by introducing attention-based architectures capable of capturing interactions across datasets between news sentiment and price dynamics [3]. Their model incorporated both LSTM components and numerical attention mechanisms, demonstrating that selectively weighting temporal features improved trend forecasting performance under noisy conditions. Importantly, their results suggested that attention mechanisms can compensate for imperfect or sparse data by amplifying informative signals. This insight directly informed our interest in evaluating whether hybrid models retain value when operating on coarse daily data rather than high frequency price streams or dense textual data. Further evidence supporting multimodal fusion comes from research proposing ensemble architectures that integrate deep sequence models with gradient boosting methods [4]. These hybrid models leverage the ability of LSTMs to capture temporal dependencies while using tree models to learn nonlinear interactions and feature importance. The study showed that stacking models can yield substantial gains across multiple evaluation metrics, reinforcing the broader trend that hybridization, both within neural networks and across model families often outperforms standalone approaches. Their findings strengthened the core question guiding our research of whether these hybrid gains persist when data availability resembles the constraints faced by everyday retail investors. Finally, the study most closely aligned with our project and really our baseline, is the FinBERT-LSTM paper. This study integrates FinBERT embeddings with sequential LSTM layers to predict short term stock movements [5]. Their results showed that the FinBERT-LSTM consistently outperformed LSTM and DNN models across multiple metrics, confirming the value of combining numerical and textual signals. However, their dataset that was over 843,000 Benzinga articles paired with NASDAQ-100 pricing data offered a dramatically richer textual environment than what typical investors encounter. Despite this, the study provided strong support for our choice of FinBERT and for employing a hybrid architecture, while simultaneously highlighting the lack of research on whether such benefits persist under low data conditions. This gap directly motivates our investigation.

3 Methodology

The goal of our methodology was to create a realistic and fair scenario to test our hybrid model. Our intention was to see if the combination of price data with news sentiments helped in any way even when the data was not perfect or large. This is a common situation the everyday investor would face.

3.1 Data Preparation

We used two main data sources, focusing on daily data as that is most easily accessible. These sources were Stock Price Data and News Text Data.

Daily stock data for Apple (AAPL) was collected from Yahoo Finance. This data covered a span of around eight years (2008 to 2016). From the closing price we were able to calculate four simple key indicators:

- Moving Averages (MA5, MA10): These were the average closing prices over the last 5 and 10 days. They display the short-term trend.
- Volatility: This measures how much the price would move each day over the last 10 days. With this, we could determine how risky or unstable a stock has been
- Momentum: The change in price of a stock in the past 10 days, giving us insight into whether the stock is accelerating up or down

Our target was a fairly simple prediction: Did the stock go up or down the next day? (1 for up and 0 for down)

For our News Text Data we gathered news headlines from a public dataset of the Top 25 headlines from Reddit's World News channel. The 25 headlines for a single day would be combined into one large block of text. This would represent the single, low-resolution block of news the every day investor might see, instead of a continuous feed.

The news text was converted into a numerical format so that it could be used in our model. We used FinBERT, which is a specialized artificial intelligence model trained to understand financial language. For each day's news block this model would provide us with a single large numerical vector (768 numbers) which was called the CLS embedding. The vector allowed for the capture of the overall sentiment and financial tone of the news for that specific day.

3.2 Data Fusion and Sequence Creation

Finally we merged the two data types:

1. Numerical Price Indicators (MA5, Volatility, etc.)
2. News Text Embeddings (the 768-dimension vector from FinBERT)

The numerical data was scaled so the values would be comparable to one another. This was done in an attempt to make sure that a single feature wouldn't overpower another.

For our advanced models, such as LSTMs, we had to look into history and not just consider a single day at a time. To begin this we converted the daily data we had gathered into a series of overlapping 30-day sequences. This allowed the LSTM to learn temporal patterns, essentially breaking down how an individual stock's behavior over the course of a month, would influence the next day's movement.

3.3 LSTM Model Architecture (The Hybrid Model)

The deep learning model we used for this project was a Long Short-Term Memory (LSTM) network. LSTM's are uniquely capable in handling sequences such as time like in our case. This model is also great for remembering significant information over a given sequence.

- LSTM Layer: The network takes the 30-day history and processes it
  - The Hybrid LSTM is the only model that would process the full data: both the price indicators and the 768 news embedding values at each of the 30 steps
- Prediction: After the 30 days of processing the LSTM provides output which is fed into a final classification layer. This simply predicts the next day's movement
- Training: The model was trained to minimize the error in prediction using BCEWithLogitsLoss over 10 rounds (epochs)

4 Experimental Design

We designed our experiment to directly look into the power of the hybrid approach and how it stacks up against the simpler, price-only methods.

4.1 Comparison Models

We looked into four different models to assess the impact of each choice:

Table 1: Models

Model Type	Data Used	Purpose
Logistic Regression (LR)	Price indicators only	The simplest baseline model.
Random Forest (RF)	Price indicators only	A more powerful, non-linear baseline model.
Price-only LSTM	30-day price sequences only	A sequence model showing whether history alone helps.
Hybrid LSTM	30-day price + news sequences	Tests whether combining modalities improves prediction under

4.2 Rigorous Testing Setup

We utilized a strict, time-respecting split since stock data is time-dependent.

- Training: The oldest 80% of the data was used to teach the models
- Testing: The most recent 20% of the data was used for final performance testing. The models were never allowed to see this testing data during training, which ensures an honest, real-world evaluation

The LSTM models were trained with a sequence length of 30 days and 10 epochs.

4.3 Performance Metrics

To evaluate how well the models predicted the next day's price movement, we used three standard metrics:

- Accuracy: the percentage of all correct predictions
- F1 Score: This was important for balancing accuracy with precision
- AUC (Area Under the Curve): This metric measures the model's overall ability to distinguish between "up" and "down" movements

Comparing the Price-Only and the Hybrid LSTMs was the most significant part of our design. This is what showed the added value of using the news sentiment when all of the other conditions are kept the same.

5 Results and Analysis

In this section, we evaluate the performance of four models, two classical baselines Logistic Regression and Random Forest, a price only LSTM, and our Hybrid LSTM integrating FinBERT news embeddings.

5.1 Baseline Models

We first assess two classical baselines trained solely on engineered price indicators (MA5, MA10, Volatility, Momentum). Across all metrics, both models struggled to extract predictive signal from daily data. Logistic Regression achieved an accuracy of **0.511**, F1 score of **0.500**, and AUC of **0.511**, while Random Forest produced slightly higher F1 score of **0.601** but the lowest accuracy overall with **0.467**. These results are consistent with prior findings that price indicator models reach a performance ceiling in short term financial prediction.

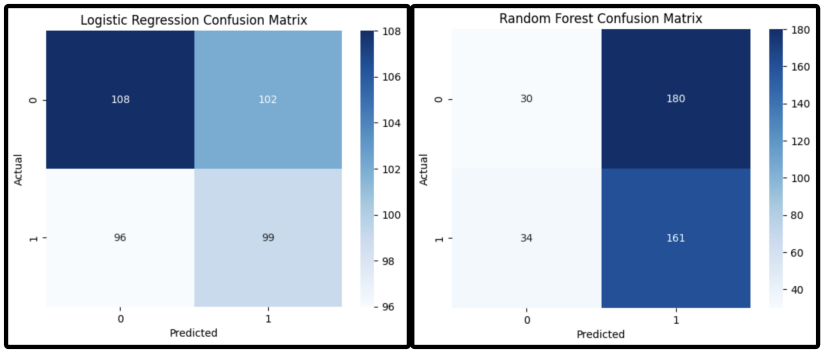


Figure 2: **Left:** Confusion Matrix for Logistic Regression. **Right:** Confusion Matrix for Random Forest. The confusion matrices for both models highlight a tendency to repeatedly misclassify upward vs downward movements, reinforcing the difficulty of using low resolution numerical signals alone.

5.2 Price Data LSTM

We next evaluate an LSTM trained on 30-day sequences of the four engineered price indicators. Despite leveraging temporal structure, the price only LSTM achieved an accuracy of **0.464**, F1 score of **0.611**, and AUC of **0.473**. While the F1 score is comparable to the baseline models, both accuracy and AUC remain low. This indicates that daily price sequences provide limited predictive signal for next day movement and that sequence modeling alone does not significantly outperform shallow baselines. To further examine model behavior, we visualize the training loss curves for both the price only and hybrid LSTM models. The price only LSTM exhibits higher and less stable loss across epochs, reinforcing the conclusion that price data alone lacks the structure needed for effective temporal learning.

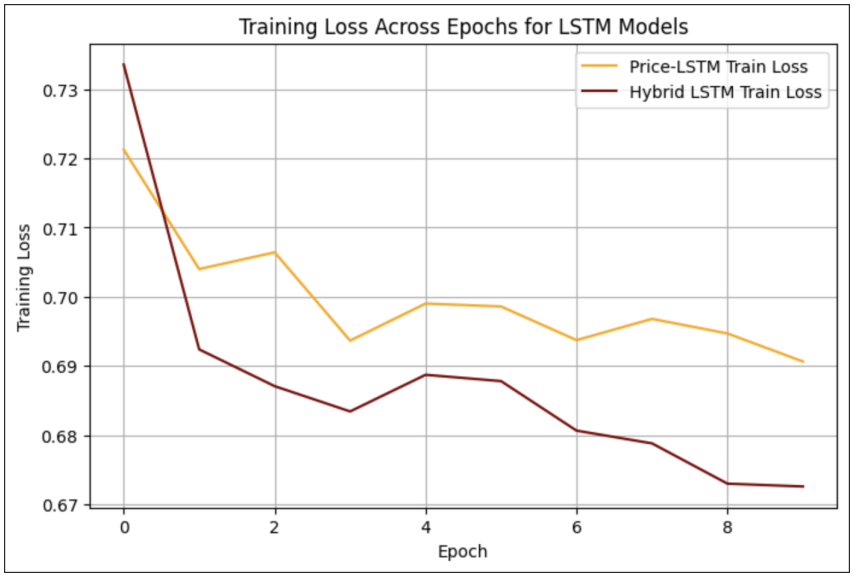


Figure 3: Training loss across epochs for the price LSTM and Hybrid LSTM models. While both curves fluctuate due to the small dataset size, the Hybrid LSTM demonstrates a more consistent downward trend and converges to a lower final loss, indicating that sentiment embeddings provide additional predictive signal beyond what is available in daily price movements alone.

5.3 Hybrid LSTM

The Hybrid LSTM, which combines each day's numerical indicators with a FinBERT CLS embedding derived from aggregated Reddit world news headlines, delivers the strongest results across all metrics. The Hybrid model achieved:

- **Accuracy:** 0.549
- **F1 Score:** 0.628
- **AUC:** 0.559

These improvements demonstrate that sentiment provides meaningful, complementary signal even in a low data environment with not very detailed data. The nearly 0.10 increase in AUC relative to the price only LSTM indicates stronger discriminative ability in distinguishing upward vs downward days.

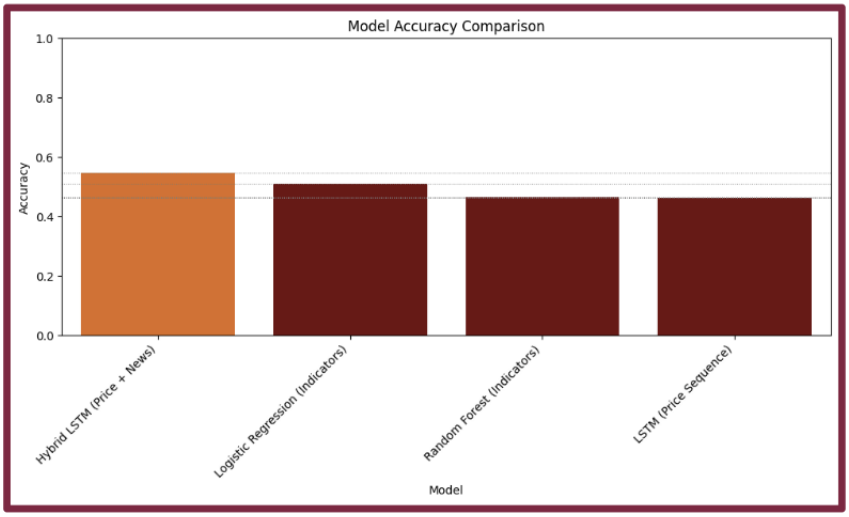


Figure 4: Bar chart showing the test accuracies of all models, with the hybrid LSTM achieving the highest performance.

5.4 Model Comparison and Key Insights

This table summarizes the performance of all four models in a numerical fashion:

Model	Accuracy	F1 Score	AUC
Logistic Regression (Indicators)	0.511	0.500	0.511
Random Forest (Indicators)	0.467	0.601	0.480
LSTM (Price Sequence)	0.464	0.611	0.473
Hybrid LSTM (Price + News)	**0.549**	**0.628**	**0.559**

Table 2: Final model comparison across all evaluation metrics.

Several key insights emerge:

- **Hybridization matters more than sequence modeling alone.** The LSTM becomes competitive only when textual sentiment is incorporated.
- **AUC gains are especially meaningful** in financial contexts where even small improvements in directional discrimination drive practical value.
- **Unexpected behavior**: Random Forest and the price LSTM achieve F1 scores above 0.60 but poor accuracy, suggesting systematic overprediction of the majority class.

- **Primary conclusion**: Even under low data, low resolution conditions, multimodal models have an advantage.

6 Discussion

The results of our study highlight several important insights regarding the behavior and practical utility of hybrid deep learning models for short-term stock prediction under realistic low-data conditions. Across all models evaluated, we consistently observed that the Hybrid LSTM consistently achieved the strongest performance in accuracy, F1 score, and AUC. This finding supports our central hypothesis: combining numerical price indicators with textual sentiment features yields measurable predictive gains even when both data sources are limited in resolution and scope. Notably, the hybrid model outperforms not only the classical baselines such as Logistic Regression and Random Forest, but also the price-only LSTM, despite sharing the same sequence architecture. This performance gap suggests that the added information from daily sentiment embeddings allows the network to capture aspects of market behavior that price dynamics alone do not represent.

A key theme emerging from our analysis is accessibility. Many stock prediction studies operate under assumptions that do not reflect the constraints faced by everyday investors, such as access to minute level price feeds or proprietary financial news streams. By contrast, all components of our dataset including Yahoo Finance daily indicators and Reddit world news headlines are freely available and easy to gather. The fact that the Hybrid LSTM delivers superior performance within this constrained environment underscores its practicality outside of institutional settings. Our results suggest that hybrid modeling can meaningfully benefit users who rely on public data sources, including students, hobbyist traders, or small research groups.

Another notable observation is that multi-modal learning remains effective even when each modality is individually limited. Our sentiment signal is derived from aggregated daily headlines, producing a single text block per day, offering only a coarse representation of market sentiment. Despite this limitation, FinBERT-derived embeddings significantly improved accuracy when incorporated alongside price sequences. This reinforces the idea that sentiment captures forms of information such as geopolitical shifts, macroeconomic signals, or a public reaction that often precedes or amplifies market movements. Our results show that even low-resolution sentiment appears valuable and provides an actionable signal when fused with numerical features.

However, our findings must be interpreted within the limitations of the study design. First, our data set focuses exclusively on a single asset (AAPL), and the behavior of hybrid models can vary between stocks with lower liquidity or greater volatility. Second, our numerical indicators were intentionally minimal, reflecting common features accessible to everyday users but excluding technical indicators or volume-based signals. Although this design choice isolates the effect of sentiment, it likely under represents the full space of price-based information. Finally, the coarse daily aggregation of Reddit headlines, while realistic for non-professional users, may omit meaningful intraday events that influence real market movement. These limitations suggest several directions for future work. Incorporating more granular news data or intraday price series could help evaluate whether hybrid benefits persist or even increase when richer data is available. Similarly, exploring Transformer based prediction heads or cross asset modeling may uncover additional structure that LSTMs alone cannot capture. Nevertheless, even under our intentionally constrained conditions, the Hybrid LSTM demonstrated clear, measurable gains emphasizing the practical relevance of multi-modal learning in everyday financial prediction.

In general, our findings contribute to a broader understanding of accessible stock prediction methods. They demonstrate that publicly available data, when processed with domain-tuned language models, can helps us to have a predictive performance in short-term forecasting tasks.

References

[1] Zheng, Y. & Jin, X. (2020) Stock Movement Prediction Using News and Price Data. CS230 Deep Learning Final Project Report, Stanford University.

[2] Haque, T., Mandal, S. & Uddin, M. (2018) Stock Market Prediction Using Twitter Sentiment Analysis and Machine Learning. Florida International University Technical Report.

[3] Reddy, K. & Singh, A. (2021) Attention-Based Hybrid Models for Stock Market Prediction Using Technical Indicators and Financial Sentiment. International Institute of Information Technology (IIIT) Research Report.

[4] Yu, C., Liu, F., Zhu, J., Guo, S., Gao, Y., Yang, Z., & Liu, M. (2025). Gradient Boosting Decision Tree with LSTM for Investment Prediction. arXiv:2505.23084.

[5] Gu, S., Kelly, P. & Xie, R. (2024) FinBERT-LSTM for Stock Movement Prediction: Integrating Financial Text Embeddings with Sequential Deep Learning. Benzinga–NASDAQ Financial Text Research Study.

[7] Aroussi, R. (2024). yfinance: Yahoo! Finance market data downloader. Official documentation: https://ranaroussi.github.io/yfinance/

[8] Sun, A. (2016). Stock News Dataset. Kaggle. https://www.kaggle.com/datasets/aaron7sun/stocknews

[9] Araci, D. (2019). FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. Available from https://huggingface.co/ProsusAI/finbert

A Technical Appendices and Supplementary Material

A.1 GitHub Repository

`https://github.com/aadisrivastava2023F/VTCS4824MachineLearningCapstoneStockPrediction`

This repository includes:

- Our completed final .ipynb notebook
- **README** describing how to run code
- Requirements file
- Data preprocessing and feature generation scripts
- Entire model creation and training

A.2 Additional Experimental Figures

The following section include plots used for analysis, but beyond not shown in the main paper.

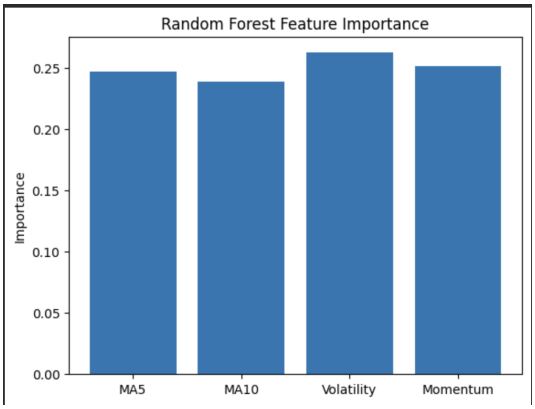


Figure 5: Bar chart showing the Random Forest model's feature importances across MA5, MA10, volatility, and momentum.

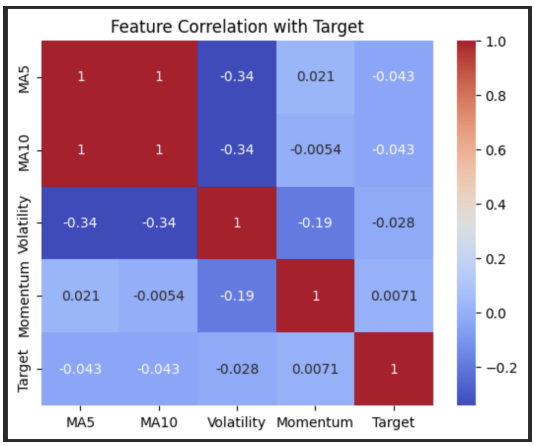


Figure 6: Correlation heatmap showing the relationships between all features and the target, highlighting that price based indicators have weak linear correlation with next day movement.

A.3 Use of AI Tools

We used OpenAI ChatGPT and Google Gemini as a writing and debugging assistant during the preparation of this report.