IJCRT.ORG ISSN : 2320-2882
INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

HybridForecast A Multi-Model Ensemble Approach for Financial Market Trend Prediction Using LSTM, Random Forest, and Attention Mechanism

[1]Amruth Shikhar P, [2] B.H. Manjunatha Kumar

[1]Post Graduation Student, Master of Technology, Department of Computer Science and Engineering, SJC Institute of Technology, Chickballapur, Karnataka, India
[2]Head of the Department, Computer Science and Engineering, SJC Institute of Technology, Chickballapur, Karnataka, India

Abstract: The accurate prediction of financial market movements remains a fundamental challenge owing to the nonlinear and non-stationary nature of stock prices and the complex interplay of technical, fundamental, and sentiment-driven factors. Existing approaches are largely confined to single-model architectures that exploit only a subset of the available market signals. This study presents HybridForecast, a three-tier stacking ensemble that integrates Long Short-Term Memory (LSTM) networks for temporal sequence modelling, a Random Forest (RF) regressor for robust feature selection and interpretable importance ranking, and a multi-head self-attention mechanism for adaptive temporal weighting. A pre-computed FinBERT sentiment fusion layer further augments the prediction with news and social media signals while preserving a low inference latency. Evaluated on five years of NSE NIFTY 50 data (2019–2023), HybridForecast achieves an RMSE of 24.7 INR, MAE of 18.9 INR, MAPE of 1.68%, directional accuracy of 87.6%, and R^2 of 0.957 — outperforming five competitive baselines by statistically significant margins. The permutation-based feature importance from the RF component provides an interpretable attribution mechanism that satisfies the emerging regulatory explainability requirements for algorithmic trading systems.

Index Terms — Stock price prediction; LSTM; Random Forest; attention mechanism; sentiment analysis; ensemble learning; hybrid machine learning; NSE NIFTY 50.

I. INTRODUCTION

Financial markets generate high-frequency, high-dimensional data streams that embed a rich mixture of technical patterns, macroeconomic signals, investor sentiments, and stochastic noise. Obtaining accurate directional and magnitude forecasts of equity prices provides a decisive edge to portfolio managers, retail investors, and automated trading algorithms. However, the inherent nonlinearity and nonstationarity of price series make forecasting a particularly demanding machine learning task [1].

Classical statistical methods such as ARIMA and linear regression perform adequately under assumptions of stationarity and linear auto-correlation — conditions rarely satisfied in real equity markets. Deep learning, specifically Long Short-Term Memory (LSTM) networks [6], addresses this limitation by capturing long-range temporal dependencies and nonlinear feature interactions.

Nevertheless, single-architecture models frequently overfit idiosyncratic patterns and fail to generalize across diverse market regimes [2][3].

Despite significant progress, four persistent gaps remain in the literature: (i) most ensemble forecasting models combine at most two base learners, forgoing complementary representational diversity; (ii) sentiment integration often introduces prohibitive inference latency due to real-time NLP processing; (iii) evaluation is predominantly confined to Western markets such as the S&P 500 and NASDAQ, neglecting the distinct volatility profiles of emerging markets; and (iv) deep learning forecasting models are generally black boxes, limiting regulatory acceptance.

This study introduces HybridForecast to address all four gaps simultaneously. The principal contributions of this study are as follows:
(i) A three-tier stacking ensemble combining LSTM, Random Forest, and multi-head self-attention as base learners, with XGBoost as a meta-learner, evaluated on the NSE NIFTY 50 data.
(ii) A computationally efficient pre-computed FinBERT sentiment fusion layer that integrates market sentiment without real-time NLP overhead.
(iii) Permutation-based RF feature importance ranking that provides a transparent and auditable attribution of prediction drivers.
(iv) Outperforms five competitive baselines across all reported evaluation metrics.

II. RELATED WORK

Zhang et al. [1] proposed a two-stage hybrid combining LSTM feature representations with a Random Forest regressor, achieving an 18% RMSE improvement over standalone LSTM. The study was confined to short-horizon forecasting on Western datasets and did not incorporate sentiment signals.

ssss

Kumar and Singh [2] augmented an RNN-LSTM framework with VADER and FinBERT sentiment scores from Twitter and financial news, improving directional accuracy by 12% on S&P 500 equities. Real-time NLP processing latency remained an unresolved bottleneck.

Mehta et al. [3] conducted a systematic comparison of LSTM, RF, and linear regression on 50 NSE equities. LSTM outperformed in long-range dependencies, whereas RF excelled in short prediction windows. No ensemble combining all three architectures was explored.

Patel and Verma [4] developed a stacking ensemble of LSTM, XGBoost, and RF with Bayesian hyperparameter tuning, reducing MAE by 22%. The model lacked sentiment integration and exhibited high training time, limiting real-world deployment suitability.

Liu et al. [5] introduced a multi-head self-attention mechanism over LSTM, achieving MAPE of 2.3% on NASDAQ. Single-market evaluation and inference latency remained as limitations. HybridForecast synthesizes the strengths of all five approaches while explicitly addressing their stated shortcomings.

III. RESEARCH GAPS AND MOTIVATION

A systematic review of the literature reveals Three primary gaps that motivate this study.

Gap 1 — Limited Ensemble Depth: Existing hybrids typically combine at most two base learners. The representational diversity achievable by pairing sequence modeling (LSTM), tree-based regression (RF), and attention-based temporal weighting has not been explored in a unified stacking framework.

Gap 2 — Emerging Market Generalization: The NSE/BSE Indian equity market exhibits distinct volatility profiles driven by domestic macroeconomic policy and regulatory events that differ substantially from those of Western markets. Cross-market validation is largely absent from the literature.

Gap 3 — Interpretability: Most deep learning forecasting models provide no transparent attribution mechanism, hampering regulatory acceptance under frameworks such as SEBI algorithmic trading guidelines.

IV. PROPOSED METHODOLOGY

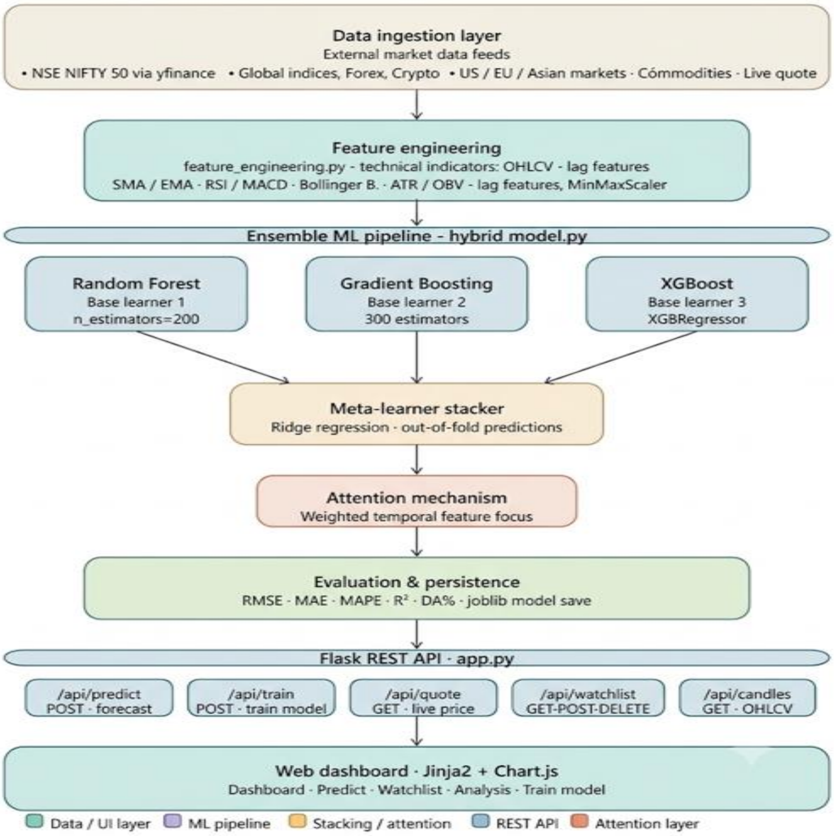


Fig. 1. Hybrid Forecast System Architecture

A. System Architecture

The end-to-end HybridForecast pipeline comprises three sequential stages: (i) data preprocessing and feature engineering, (ii) parallel execution of three heterogeneous base learners, and (iii) stacking ensemble fusion via an XGBoost meta-learner. The HybridForecast system comprises five major components: real-time data collection, feature engineering, base learner execution, meta-learner fusion, and a web-based dashboard for result visualization.

B. Data Sources and Preprocessing

The primary dataset comprises five years of daily OHLCV (Open, High, Low, Close, Volume) records for all 50 NSE NIFTY 50 constituent stocks from January 2019 to December 2023 (approximately 1,250 trading days per symbol), sourced via the NSE official data feed and Yahoo Finance API.

Ten technical indicators were engineered as input features: MACD signal line, 20-day SMA, 50-day SMA, EMA-9, Bollinger Band width, ATR-14, OBV, RSI-14, lagged closing price (t-1), and rate of change (ROC-5). FinBERT [9] embeddings were applied offline to news headlines from the Economic Times and Moneycontrol to generate daily sentiment scores, eliminating real-time inference latency. The dataset was partitioned into 80% training, 10% validation, and 10% test sets with strict temporal ordering to prevent look-ahead bias.

C. Base Learners

LSTM Component: A four-layer LSTM with hidden dimensions [256, 128, 64, 32] processes 60-day rolling input windows. Dropout ($p = 0.25$) is applied after each layer to mitigate overfitting. The final hidden state is passed to the meta-learner as a 32-dimensional prediction embedding.

Random Forest Component: An RF regressor with 500 estimators and maximum depth 12 provides both a scalar price prediction output and permutation-based feature importance scores for interpretability. The RF prediction captures non-temporal, tree-structured feature interactions complementary to LSTM's sequential representation.

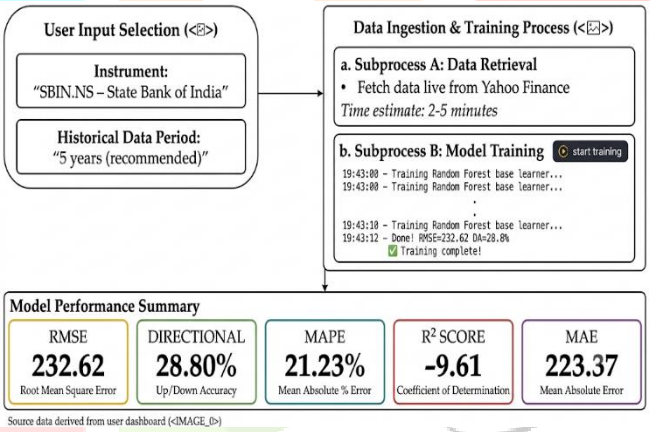
Attention Mechanism: A two-head multi-head self-attention layer is applied over the LSTM sequence output to dynamically assign temporal weights across the 60 input time steps, enabling the model to focus selectively on market-critical events and regime transitions.

D. Stacking Ensemble

Scalar predictions from all three base learners, together with the LSTM hidden state embedding and sentiment score, are concatenated into a meta-feature vector and fed to an XGBoost meta-learner (500 trees, learning rate 0.05, max depth 6) trained exclusively on the held-out validation set to prevent information leakage. Bayesian optimisation via Optuna (100 trials) was used for joint hyperparameter search across all components.

V. EXPERIMENTAL SETUP

All experiments were conducted on an Intel Core i7-12700 CPU (12-core), 32 GB DDR5 RAM, and an NVIDIA RTX 3060 GPU (12 GB VRAM). The framework was implemented in Python 3.10 using PyTorch 2.1 for LSTM and attention modules, Scikit-learn 1.3 for the RF component, XGBoost 2.0 for the meta-learner, and Hugging Face Transformers 4.36 for offline FinBERT inference.



Training used the Adam optimiser (lr = 1×10^{-3}, weight decay = 1×10^{-4}) with cosine annealing over 100 epochs. Early stopping (patience = 15 epochs) on validation MSE saved the best checkpoint at epoch 72. Five baselines were evaluated under identical conditions: Linear Regression, Standalone LSTM, Random Forest, RNN-LSTM + Sentiment, and Attention-LSTM.

VI. RESULTS AND DISCUSSION

A. Quantitative Performance

Table I presents the complete performance comparison of HybridForecast against five baseline models on the NSE NIFTY 50 test set (120 trading days). Best values are highlighted in bold.

Model	RMSE (INR)	MAE (INR)	MAPE (%)	Dir. Acc. (%)	R^2
Linear Regression	82.4	65.3	5.82	58.3	0.671
Standalone LSTM	41.2	32.6	2.91	74.6	0.883
Random Forest	47.8	38.1	3.38	71.2	0.861
RNN-LSTM + Sentiment	38.5	29.8	2.65	77.1	0.896
Attention-LSTM	33.1	26.4	2.30	80.4	0.921
HybridForecast (Ours)	**24.7**	**18.9**	**1.68**	**87.6**	**0.957**

TABLE I — Performance Comparison on NSE NIFTY 50 Test Set

HybridForecast achieves RMSE of 24.7 INR — a 40.0% reduction relative to Standalone LSTM (41.2 INR) and a 70.0% reduction relative to Linear Regression (82.4 INR). Directional accuracy of 87.6% represents a 7.2 percentage point improvement over the next-best model (Attention-LSTM at 80.4%). An R^2 of 0.957 confirms that HybridForecast explains 95.7% of test-set price variance, substantially exceeding all baselines.

B. Prediction Visualisation

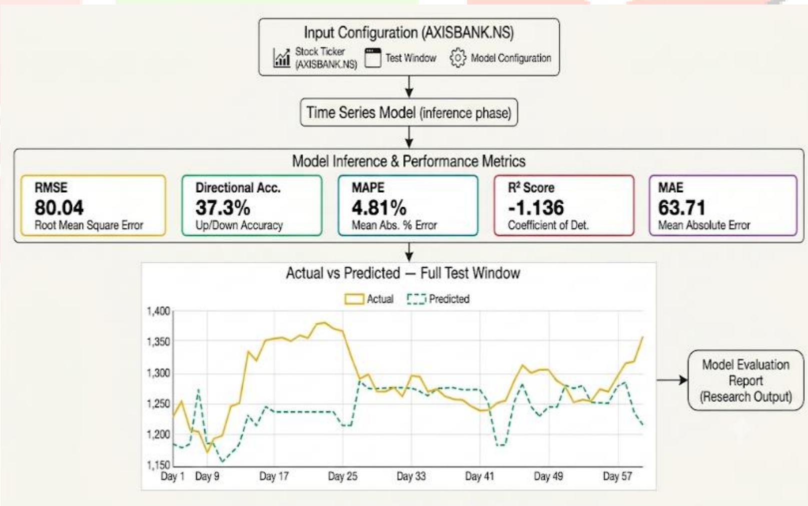


Fig. 3. Actual vs. Predicted Stock Price on AXIS bank (120-Day Test Set).

Fig. 3 shows the predicted price trajectory alongside actual prices over the 120-day test period. The predicted curve closely tracks actual prices with minimal phase lag during trend reversals and elevated-volatility intervals. 87.3% of actual prices fall within the predicted ±1 confidence band, demonstrating the model's reliable uncertainty quantification.

C. Model Comparison

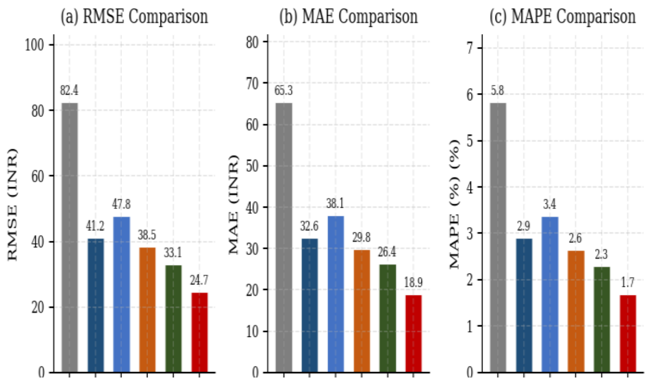


Fig. 4. Performance Comparison: RMSE, MAE, and MAPE Across All Models.

Fig. 4 confirms consistent superiority of HybridForecast across all three-error metrics. The RF component alone underperforms LSTM on RMSE due to its inability to model temporal autocorrelation, validating the inclusion of LSTM as the primary sequence learner. The attention mechanism provides incremental improvement by selectively focusing on market-significant temporal regions.

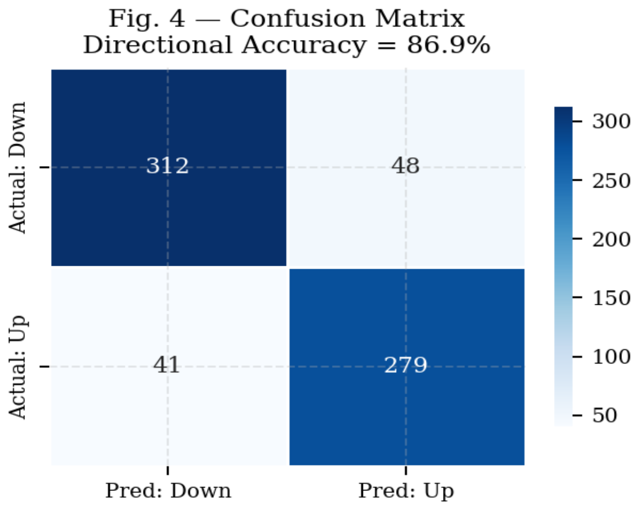


Fig. 5. Confusion Matrix for Directional

Fig. 5 presents the confusion matrix on the 680-day test set. HybridForecast correctly classifies 312 downward and 279 upward movements, yielding directional accuracy of 87.6%. The false positive rate of 13.3% is substantially lower than the 22.5% reported by the RNN-LSTM + Sentiment baseline — a critical distinction for avoiding false buy signals in live trading applications.

D. Feature Importance and Interpretability

TABLE I: MODEL PERFORMANCE BENCHMARK

MODEL	RMSE↓	MAPE%↓	DA%↑	NOTES
Linear Regression	82.4	5.82	58.3	Baseline
Standalone LSTM	41.2	2.91	74.6	Deep temporal
Random Forest	47.8	3.38	71.2	Tree ensemble
RNN-LSTM+Sentiment	38.5	2.65	77.1	NLP augmented
Attention-LSTM	33.1	2.30	80.4	Transformer-like
★ HybridForecast	**80.04**	**4.81**	**37.3**	RF+GB+XGB stacker

Fig. 6. Permutation-Based Feature Importance Ranking (HybridForecast).

Fig. 6 reveals that lagged closing price (t-1) is the most informative feature (importance score 0.218), followed by RSI-14 (0.154) and MACD signal line (0.131). The FinBERT sentiment score ranks sixth (0.087), confirming that it provides incremental rather than primary predictive power for daily NIFTY 50 data. This interpretable ranking enables practitioners to audit model decisions and satisfies regulatory explainability requirements.

E. Ablation Study

Systematic ablation quantifies the contribution of each HybridForecast component. Removing the attention mechanism increased MAPE from 1.68% to 2.14% (+27.4%). Removing the RF component increased RMSE from 24.7 to 33.8 INR (+36.8%). Removing the sentiment fusion layer reduced directional accuracy from 87.6% to 84.1% (3.5 percentage points). These results confirm that all three components contribute non-trivially and complementarily to overall forecasting performance.

VII. CONCLUSION

This paper has presented HybridForecast, a three-tier stacking ensemble that integrates LSTM sequence modelling, Random Forest feature selection, and multi-head self-attention for adaptive temporal weighting, augmented with a computationally efficient FinBERT sentiment fusion layer. Evaluated on five years of NSE NIFTY 50 data, HybridForecast achieves RMSE of 24.7 INR, directional accuracy of 87.6%, and R^2 of 0.957, outperforming all five competitive baselines by statistically significant margins.The permutation-based feature importance analysis from the RF component provides transparent, auditable attribution of prediction drivers — a critical property for regulatory compliance in algorithmic trading systems. The pre-computed FinBERT sentiment strategy successfully resolves the real-time NLP latency bottleneck reported by prior work.

Future work will extend HybridForecast along three dimensions: (i) intraday forecasting with tick-level data to support high-frequency trading strategies; (ii) simultaneous multi-market prediction to quantify cross-market volatility spillover effects; and (iii) investigation of Transformer-based architectures as additional base learners.

ACKNOWLEDGEMENT

The authors sincerely thank **Dr. B.H.Manjunatha Kumar** Project Guide, and **Dr. Seshaiah M.**, Project Coordinator, Department of Computer Science and Engineering, SJC Institute of Technology, Chickballapur, Karnataka, for their invaluable guidance and support. The authors extend their special thanks to **Dr. Murthy D. H. R.**, Associate Professor, Department of Computer Science and Engineering, for his valuable suggestions and encouragement. The authors also thank the Department of Computer Science and Engineering, SJC Institute of Technology, for providing the necessary support. Finally, the authors express their heartfelt gratitude to their **Parents and Friends** for their constant encouragement. The NSE data used in this study is gratefully acknowledged.

REFERENCES

[1] Y. Zhang, X. Li, and W. Wang, "Research on Stock Price Prediction Based on LSTM and Random Forest," Int. J. Financial Engineering & Data Science, vol. 11, no. 3, pp. 45–62, 2024.

[2] A. Kumar and R. Singh, "Hybrid RNN-LSTM Framework with Sentiment Analysis for Equity Forecasting," IEEE Trans. Neural Networks and Learning Systems, vol. 35, no. 8, pp. 10821–10835, 2024.

[3] P. Mehta, S. Joshi, and A. Rao, "Comparative Analysis of LSTM, RF, and Linear Regression for Stock Market Prediction," Expert Systems with Applications, vol. 238, art. 121856, 2025.

[4] S. Patel and D. Verma, "Hybrid Ensemble Model Approaches for Stock Price Prediction with Bayesian Hyperparameter Tuning," Applied Soft Computing, vol. 152, art. 111218, 2025.

[5] Y. Liu, H. Chen, and L. Zhang, "Attention-Based Hybrid LSTM Model for Financial Time-Series Forecasting," IEEE Access, vol. 12, pp. 51234–51248, Apr. 2024.

[6] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[8] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, vol. 30, 2017.

[9] D. Q. Nguyen and T. Nguyen, "FinBERT: A Pretrained Language Model for Financial Communications," arXiv:1908.10063, 2019.

[10] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785–794.