International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Apex Ai:
A Multi-Model Ensemble Framework for Intelligent NSE Equity Trading Signal Generation

Sai Narendra Ghodke, Siddhartha V. Bhosale, Sunraj Shetty
Dept. of Artificial Intelligence and Machine Learning Rasiklal M. Dhariwal Institute of Technology, Pune, India

Abstract- — This paper presents APEX AI, a professional-grade equity trading signal platform designed for National Stock Exchange (NSE) listed Indian stocks. The system employs a heterogeneous ensemble of three complementary machine learning models: Gated Recurrent Unit (GRU) networks for sequential pattern capture, Temporal Convolutional Networks (TCN) for multi-scale temporal feature extraction, and LightGBM for gradient-boosted tabular learning. These models are fused through a soft-voting ensemble to produce probabilistic price forecasts expressed as P10, P50, and P90 quantile estimates over a 14-day horizon. A four-stage gate architecture governs signal quality, filtering signals based on trend alignment, volatility regime, volume confirmation, and risk-adjusted expected return. The platform exposes predictions through a FastAPI backend and a React/TypeScript/Vite frontend featuring a TradingView-style candlestick chart with an integrated forecast cone. Experimental evaluation on historical NSE data demonstrates directional accuracy above 62%, with the ensemble outperforming any individual constituent model.

Keywords — Deep Learning, Gated Recurrent Unit, Temporal Convolutional Network, LightGBM, Ensemble Learning, Stock Market Prediction, NSE, Trading Signals, Quantile Forecasting, FastAPI.

I. INTRODUCTION

The Indian equity market, anchored by the National Stock Exchange (NSE), is among the fastest-growing financial markets globally. Retail participation has surged in the post-COVID era, yet most individual investors rely on intuition, rudimentary chart analysis, or lagged brokerage advisories. The absence of systematic, data-driven signal generation tools tailored to Indian market microstructure represents a significant gap.

Classical technical analysis (moving averages, RSI, MACD) suffers from parameter sensitivity and overfitting in non-stationary markets. Statistical time series models such as ARIMA and GARCH require stationarity assumptions that equity price series routinely violate. Recent advances in deep learning, particularly recurrent architectures and gradient-boosted trees, have demonstrated superior performance on financial time series owing to their capacity to capture nonlinear dependencies and long-range temporal patterns [1][2].

APEX AI addresses this gap by combining the complementary strengths of three model families into a unified, probabilistic forecasting pipeline. Rather than producing a single-point price prediction — which is both unreliable and non-actionable — the system generates a 14-day forecast cone anchored at the P10, P50, and P90 quantiles, communicating uncertainty to the trader and enabling position sizing proportional to signal confidence.

The primary contributions of this work are:

- A heterogeneous ensemble (GRU + TCN + LightGBM) with soft-vote fusion for NSE equity price forecasting.
- A probabilistic 14-day forecast expressed as P10/P50/P90 quantiles, enabling uncertainty-aware trading decisions.
- A four-stage gate architecture that filters raw model output through trend, volatility, volume, and return-quality screens before a signal is emitted.
- A full-stack trading signal platform built on FastAPI and React/TypeScript with a TradingView-style candlestick interface.

II. RELATED WORK

Stock market prediction using machine learning has been extensively studied. Early approaches relied on feed-forward networks [3] and SVMs [4], which struggled with the sequential nature of price data. LSTM networks introduced by Hochreiter and Schmidhuber [5] became the dominant deep learning approach, with numerous studies reporting improved directional accuracy over statistical baselines [6].

Gated Recurrent Units (GRU), proposed by Cho et al. [7], reduce LSTM parameter count while retaining comparable performance on sequence tasks. TCNs, introduced by Bai et al. [8], apply dilated causal convolutions to achieve large receptive fields with better parallelism than recurrent models, and have shown strong results on financial time series benchmarks.

Gradient-boosted decision trees, particularly XGBoost [9] and LightGBM [10], consistently win tabular prediction competitions and complement deep architectures by capturing feature interactions and handling engineered indicators efficiently. Ensemble methods combining heterogeneous learners — often referred to as stacked generalization [11] — systematically outperform single-model approaches by reducing variance and bias simultaneously.

Quantile regression for financial forecasting has been explored by Taylor [12] using CAViaR and more recently through deep quantile networks [13]. Our work synthesises these lines of research into an end-to-end, production-oriented platform specifically calibrated for NSE-listed equities.

III. SYSTEM ARCHITECTURE

APEX AI follows a layered microservices architecture. The system is composed of four principal layers: Data Ingestion, Model Inference, Signal Gating, and Presentation.

Data Ingestion Layer
Historical OHLCV (Open, High, Low, Close, Volume) data for NSE-listed equities is fetched via the yfinance library using NSE ticker symbols (e.g., RELIANCE.NS). The ingestion pipeline resamples daily candles, handles corporate actions, and forward-fills missing trading days. A rolling window of 60 trading days is used as the primary input sequence for deep learning models. Technical indicators — 14-day RSI, MACD(12,26,9), 20-day Bollinger Bands, ATR(14), OBV, and 20-day VWAP — are computed and appended to form a multivariate feature matrix.

Model Layer
The model layer hosts three independently trained forecasters: a GRU network, a TCN, and a LightGBM regressor. Each model is trained on MinMaxScaler-normalised price data and produces its own 14-day forecast. Inverse transformation is applied at inference time to convert normalised outputs back to rupee-denominated prices.

Ensemble Fusion
Model outputs are fused through a soft-voting ensemble with equal weights (1/3 each). The fused prediction is then passed through a quantile estimation block that derives P10, P50, and P90 estimates from the distribution of constituent model outputs and historical residuals. This provides natural uncertainty quantification without requiring explicit quantile training objectives.

Signal Gating Architecture
Before a trading signal is emitted, the fused forecast must pass four sequential gates:

- Gate 1 – Trend Gate: Validates that the P50 forecast direction is consistent with the 20-day EMA slope.
- Gate 2 – Expected Return Gate: Confirms the risk-adjusted return (P50 - entry) / ATR exceeds a minimum threshold.
- Gate 3 – Volatility Gate: Rejects signals during abnormal volatility regimes flagged by a 30-day ATR Z-score above 2.0.
- Gate 4 – Volume Gate: Requires that recent average volume is at least 1.5x the 90-day baseline, indicating institutional participation.

API and Frontend
The backend is implemented in FastAPI (Python 3.10), exposing a RESTful /predict endpoint that returns structured JSON containing OHLCV history, 14-day forecast (ohlcv and forecast arrays), gate results, position sizing, and a narrative explanation. The frontend, built with React 18, TypeScript, and Vite, renders a TradingView-style candlestick chart using the lightweight-charts library overlaid with a P10/P50/P90 forecast cone.

IV. METHODOLOGY

Gated Recurrent Unit (GRU)
The GRU model takes a (60, F) input tensor where F is the number of features (5 OHLCV + 6 technical indicators = 11). The network comprises two stacked GRU layers (128 and 64 units) with 0.2 dropout, followed by a Dense layer projecting to a 14-day output. The network is trained using the Adam optimiser with a learning rate of 0.001 and Huber loss (delta=1.0) over 100 epochs with early stopping (patience=10) on a 10% validation split.

Temporal Convolutional Network (TCN)
The TCN employs dilated causal convolutions with dilation factors [1, 2, 4, 8, 16] and 64 filters of kernel size 3, yielding a receptive field of 93 timesteps — sufficient to cover the 60-day

input window with margin. Residual connections are applied at each dilation block to mitigate vanishing gradients. Weight normalisation is applied in place of batch normalisation to maintain causality during inference.

LightGBM Regressor
LightGBM is trained on a flattened 60-day feature vector augmented with lag features (t-1, t-5, t-10 returns) and calendar features (day-of-week, month). The target is the 14-day forward return. Hyperparameters are tuned via 5-fold time-series cross-validation: num_leaves=63, learning_rate=0.05, n_estimators=500, min_child_samples=20.

Training and Data Split
All models are trained on five years of daily NSE data (2018-2022) and evaluated on an out-of-sample test set spanning 2023. A strict temporal split is enforced — no future data leaks into training. MinMaxScaler is fitted exclusively on the training partition, with the fitted scaler serialised alongside model weights for consistent inference.

Position Sizing
Position size is computed using a volatility-adjusted Kelly fraction: f = (P50 - S) / (P90 - P10), where S is the current spot price. The result is clipped to [0.05, 0.25] of portfolio value to enforce risk discipline. This rewards high-conviction signals with larger allocations while capping exposure on uncertain predictions.

V. RESULTS AND ANALYSIS

Models were evaluated on the 2023 out-of-sample test set covering 50 actively traded NSE equities across five sectors (Banking, IT, FMCG, Pharma, Energy). Directional accuracy (DA) measures whether the predicted 14-day close direction matches the realised direction.

Table I: Model Performance on NSE Test Set 2023

Model	DA (%)	MAPE (%)	sMAPE (%)
GRU	58.2	3.41	3.38
TCN	57.6	3.67	3.63
LightGBM	55.9	4.02	3.97
Ensemble	62.4	2.89	2.86

As shown in Table I, the soft-vote ensemble consistently outperforms all constituent models across all three metrics. The 4.2 percentage point improvement in directional accuracy over the best single model (GRU) is statistically significant ($p < 0.05$, McNemar's test). The ensemble's lower MAPE confirms that averaging diverse error profiles reduces systematic bias.

Gate analysis shows that the four-stage filter reduces the raw signal pool by approximately 68%, retaining only high-quality signals. Among filtered signals, directional accuracy improves to 67.1%, validating the gate architecture as an effective post-processing step. Volatility gate rejections were most frequent during the Q1 2023 banking sector turbulence, demonstrating appropriate risk-off behaviour.

Sector-wise breakdown reveals that IT and Banking sectors yield the highest directional accuracy (65.3% and 63.8% respectively). Energy and FMCG stocks exhibit more mean-reversion behaviour, where the momentum-oriented ensemble underperforms relative to the sector average (58.2% and 57.9%).

VI. CONCLUSION

This paper presented APEX AI, an end-to-end trading signal platform for NSE equities built on a heterogeneous ensemble of GRU, TCN, and LightGBM models. The system addresses key limitations of single-model approaches by combining complementary learning architectures, and goes beyond point prediction by delivering a probabilistic 14-day forecast cone (P10/P50/P90) alongside a four-stage quality gate that filters signals by trend, return, volatility, and volume criteria.

Empirical evaluation on 50 NSE equities over a full out-of-sample year demonstrates 62.4% directional accuracy for the ensemble versus 55.9–58.2% for individual models, with a post-gate signal pool achieving 67.1% directional accuracy. These results suggest that intelligent ensemble fusion combined with disciplined signal filtering can meaningfully improve the quality of machine learning-driven trading signals in the Indian equity market.

Future work will explore the incorporation of macroeconomic features (RBI policy rate changes, FII/DII flow data), sentiment signals derived from NSE news feeds, and Temporal Fusion Transformer (TFT) architecture which natively supports multi-horizon quantile forecasting. Walk-forward validation over multiple market regimes and live paper-trading backtests are planned to further validate production readiness.

Acknowledgment

The authors thank the faculty and students of the Department of Computer Engineering for their valuable feedback during the capstone project review. The authors also acknowledge the

open-source communities behind Keras, LightGBM, FastAPI, React, and lightweight-charts whose libraries form the technical foundation of this work.

REFERENCES

1. J. Jiang, "Stock Market Prediction Using Machine Learning: A Systematic Survey," Expert Systems with Applications, vol. 187, 2022.
2. M. Nabipour et al., "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data," IEEE Access, vol. 8, pp. 150199-150212, 2020.
3. E. Mizuno et al., "Application of Neural Network to Technical Analysis of Stock Market Prediction," Studies in Informatics and Control, vol. 7, no. 3, pp. 111-120, 1998.
4. W. Huang, Y. Nakamori, and S.-Y. Wang, "Forecasting stock market movement direction with support vector machine," Computers & Operations Research, vol. 32, no. 10, pp. 2513-2522, 2005.
5. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
6. N. Roondiwala, H. Patel, and S. Varma, "Predicting Stock Prices Using LSTM," International Journal of Science and Research, vol. 6, no. 4, pp. 1754-1756, 2017.
7. K. Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in Proc. EMNLP, 2014, pp. 1724-1734.
8. S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," arXiv:1803.01271, 2018.
9. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. ACM SIGKDD, 2016, pp. 785-794.
10. G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in NeurIPS, 2017, pp. 3146-3154.
11. D. H. Wolpert, "Stacked Generalization," Neural Networks, vol. 5, no. 2, pp. 241-259, 1992.
12. J. W. Taylor, "A Quantile Regression Approach to Estimating the Distribution of Multiperiod Returns," Journal of Derivatives, vol. 7, no. 1, pp. 64-78, 1999.
13. A. Rodrigues et al., "Deep Quantile Regression for Forecasting," International Journal of Forecasting, vol. 38, no. 3, 2022.

Author Profile

Sai Narendra Ghodke is the Lead AI Developer of the APEX AI project. His research interests include deep learning for time series, reinforcement learning for algorithmic trading, and full-stack ML system design.

Siddhartha V. Bhosale leads the quantitative finance aspects of APEX AI, including feature engineering, signal design, and risk-adjusted position sizing frameworks.

Sunraj Shetty is the UI/UX designer and frontend engineer responsible for the React/TypeScript interface, TradingView-style chart integration, and overall user experience design of the APEX AI platform.