IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MARKETGENIUS: AN ENSEMBLE MACHINE LEARNING FRAMEWORK FOR REAL-TIME INDIAN STOCK MARKET PREDICTION AND ANALYSIS

Intelligent Financial Forecasting Using Random Forest and Gradient Boosting with Technical Indicator Feature Engineering

1 Ms. Shriraksha A S

Prof , Dept. of Information Science & Engineering

Jain Institute of Technology, Davangere, India

2 Yogesh N

Dept. of Information Science & Engineering

Jain Institute of Technology, Davangere, India

3 Sheeba Mariyam

Dept. of Information Science & Engineering

Jain Institute of Technology, Davangere, India

4 Syed Md Abbas

Dept. of Information Science & Engineering

Jain Institute of Technology, Davangere, India

5 Varshith H D

Dept. of Information Science & Engineering

Jain Institute of Technology, Davangere, India

Abstract — Forecasting equity market movements in emerging economies such as India demands computational systems that can handle high volatility, nonlinear price dynamics, and abrupt structural breaks without losing temporal coherence. This work presents MarketGenius, an ensemble machine learning platform developed to forecast next-day closing prices and market direction for NIFTY 50 and SENSEX indices, as well as a selection of individual NSE and BSE equities. The platform ingests historical OHLCV data spanning January 2010 through December 2023 and constructs a 25-dimensional feature vector per trading session through a domain-driven engineering pipeline that incorporates the Relative Strength Index (RSI), Average Directional Index (ADX), Simple Moving Averages (SMA), Exponential Moving Averages (EMA), and Moving Average Convergence Divergence (MACD), augmented by multi-day lag features and rolling volatility estimates. A weighted combination of Random Forest and Gradient Boosting regressors, calibrated through five-fold cross-validation, attains a normalized price regression RMSE of 0.92, MAE of 0.71, and R^2 of 0.95 on held-out test data spanning 2022–2023. Direction classification on this unseen window records 50.56% accuracy with an F1-score of 0.53—a figure consistent with the efficient market hypothesis—while training-phase accuracy of 79.28% confirms that the models extract meaningful regime-level patterns from historical data. Compared against ARIMA and standalone LSTM baselines, the ensemble achieves 36.5% and 17.9% reductions in RMSE respectively. A Flask-based REST API exposes inference endpoints with sub-500 ms latency, and a React dashboard renders real-time predictions, technical indicator overlays, and sentiment signals interactively.

Index Terms *— Stock market prediction, ensemble learning, Random Forest, Gradient Boosting, technical indicators, feature engineering, NIFTY 50, SENSEX, Flask REST API, financial forecasting, RSI, MACD, machine learning.*

I. INTRODUCTION

Equity markets serve as barometers of macroeconomic health and aggregate investor sentiment. India's two principal benchmarks—the NIFTY 50 and SENSEX—together encompass sectors ranging from banking and information technology to energy, consumer goods, and pharmaceuticals, making them indispensable reference points for retail investors, institutional fund managers, algorithmic trading desks, and regulatory bodies alike. Reliable near-term forecasting of these indices holds tangible value across all of these stakeholder categories, and yet constructing robust predictive systems for them remains genuinely difficult.

Traditional time-series approaches such as the Autoregressive Integrated Moving Average (ARIMA) model impose stationarity and linearity constraints that limit their capacity to represent the momentum effects, volatility clustering, and abrupt regime changes that characterize real financial time series. Single machine learning models, while more expressive, tend to overfit when trained on noisy financial sequences without structured feature construction grounded in market domain knowledge. These limitations motivate an ensemble strategy that draws simultaneously on the variance-reduction properties of bagging and the bias-reduction capability of sequential boosting, anchored by a feature pipeline derived from established technical analysis practice.

MarketGenius addresses this gap through four coordinated contributions: (i) a domain-driven feature engineering pipeline integrating five technical indicators alongside lag and volatility features across a 13-year historical window; (ii) a validated weighted ensemble of Random Forest and Gradient Boosting that outperforms both individual component models and classical baselines on Indian market data; (iii) a production-grade Flask REST API capable of serving inference within 500 milliseconds; and (iv) an interactive React dashboard that consolidates real-time predictions, technical analysis overlays, and composite sentiment signals within a single deployable interface.

II. LITERATURE REVIEW

The application of supervised machine learning to financial time series has grown considerably in recent years, with ensemble methods and technical indicator–based feature engineering emerging as the dominant paradigm for short-horizon equity forecasting. Table I summarizes six representative studies published between 2023 and 2025, spanning multiple markets and methodological approaches.

Table I: Summary of Related Work (2023–2025)

Author / Year	Method Used	Dataset / Index	Key Result
Mohapatra et al., 2025	XGBoost + ensemble with technical indicators	Indian Banking Stocks (BSE)	Up to 98% accuracy for medium/long-term prediction
Mondal et al., 2025	LSTM + Random Forest integration	General stock market data	99% accuracy; low MSE via complementary model fusion
Jin et al., 2025	Ridge Regression + LSTM with accuracy-weighted averaging	S&P 500 short-term	Up to 83% accuracy for next-day direction
Okoh et al., 2025	Stacking ensemble (RF, XGB, LR) with SHAP analysis	S&P 500	Improved R^2 approaching 1.0; lower MAE vs. individual models
Verma et al., 2024	DWT-based feature engineering + ensemble with PSO tuning	NIFTY 50 Index	92.51% accuracy outperforming baselines
Ramakrishnan et al., 2023	Ensemble using RSI, MACD, SMA, EMA, ATR	Generic stock market data	Superior buy/sell signal classification in volatile markets

Three consistent findings emerge across this body of work. Ensemble approaches routinely outperform single-model counterparts by margins of 5–15% on standard evaluation metrics. Technical indicators—particularly RSI, MACD, and moving average variants—constitute among the most informative input features for capturing short-horizon momentum and trend reversal signals. Critically, the majority of prior studies remain confined to offline experimental settings and do not address the engineering challenges of real-time deployment, particularly within the Indian market context. MarketGenius is specifically designed to close this deployment gap while achieving regression accuracy that matches or exceeds published benchmarks.

III. METHODOLOGY

A. Data Collection and Preprocessing

Historical daily OHLCV records for NIFTY 50, SENSEX, and more than 50 individual NSE/BSE constituents were retrieved through the yfinance library, covering January 2010 through December 2023—approximately 3,500 trading sessions per symbol. This window was selected deliberately to encompass significant market disruptions, including the 2016 demonetization shock and the pandemic-driven crash of March 2020, ensuring that trained models are exposed to diverse market regimes rather than a single prolonged trend.

Preprocessing proceeded through four sequential stages. Missing observations—accounting for fewer than 1% of all records, attributable to market holidays and occasional data feed interruptions—were resolved by forward-fill interpolation to maintain temporal continuity. Outliers were identified via z-score thresholding at ±3 standard deviations and capped at the respective boundary values. All numerical features were then scaled to the unit interval [0, 1] using Min-Max normalization, eliminating scale disparity between price series and volume dimensions. Finally, records were sorted chronologically and partitioned into a training set (2010–2021) and a held-out test set (2022–2023), with the split boundary fixed in time to prevent any form of lookahead bias.

B. Feature Engineering and Technical Indicators

Raw OHLCV data alone provides an insufficient basis for capturing the momentum effects, trend persistence, and volatility clustering that characterize financial time series. The system therefore constructs a 25-dimensional feature vector per trading session through five categories of transformation:

RSI (14-period): Quantifies price momentum by computing the ratio of average gains to average losses over a rolling 14-day window, yielding an oscillator bounded within [0, 100]. Values exceeding 70 are conventionally interpreted as overbought conditions; readings below 30 indicate oversold territory.

ADX (14-period): Measures trend strength independently of direction, allowing the model to discriminate between trending and ranging market environments—an important distinction for ensemble weighting.

SMA (20- and 50-period): Provides smoothed representations of medium- and long-term price trajectories; crossovers between price and SMA levels serve as directional signals and are among the most widely tracked technical patterns.

EMA (9-, 12-, and 26-period): Exponentially weighted moving averages that respond more rapidly to recent price changes than their simple counterparts, capturing near-term momentum with greater sensitivity.

MACD (12, 26, 9): The difference between the 12-period and 26-period EMAs, coupled with a 9-period signal line. MACD crossovers are among the most empirically studied buy/sell trigger mechanisms in technical analysis.

Lag Features: Closing prices and volumes from the previous 1 through 5 trading sessions, encoding autocorrelation structure and short-term serial dependencies.

Rolling Volatility: The 20-day standard deviation of daily log returns, capturing heteroskedastic volatility clustering—a well-documented property of financial time series.

C. Ensemble Model Architecture

Two base learners were chosen for their complementary inductive biases. The Random Forest Regressor (100 estimators, max depth 10) achieves variance reduction through bootstrap aggregation, averaging the outputs of many decorrelated decision trees trained on random feature subsets. The Gradient Boosting Regressor (100 estimators, learning rate 0.1) reduces bias by iteratively fitting new trees to the residuals of the current ensemble. Their predictions are combined as a weighted average:

$$\hat{y} = 0.6 \times \hat{y}_{RF} + 0.4 \times \hat{y}_{GB} \quad (1)$$

Ensemble weights of 0.6 for Random Forest and 0.4 for Gradient Boosting were determined through grid search over held-out validation RMSE across five stratified folds. For direction classification, the continuous ensemble prediction is thresholded at zero implied return: a positive predicted return maps to the "Up" class, and a non-positive return maps to "Down." Hyperparameter tuning employed RandomizedSearchCV across 50 iterations, balancing exploration and computational cost.

D. System Architecture

The platform is organized across six functional layers as illustrated in Fig. 1. The Data Layer fetches raw OHLCV records from Yahoo Finance via the yfinance API and persists them in a SQLite database with scheduled refresh cycles. The Processing Layer applies the four-stage cleaning protocol described above. The Feature Engineering Layer computes all 25 technical features using TA-Lib and pandas. The Model Layer hosts the serialized Random Forest, Gradient Boosting, and weighted ensemble objects. The Prediction Layer handles regression and classification inference, including confidence interval estimation. The API and Interface Layer exposes three primary REST endpoints—/fetch_data, /symbols, and /predict—secured via JWT authentication, and serves the React dashboard frontend.

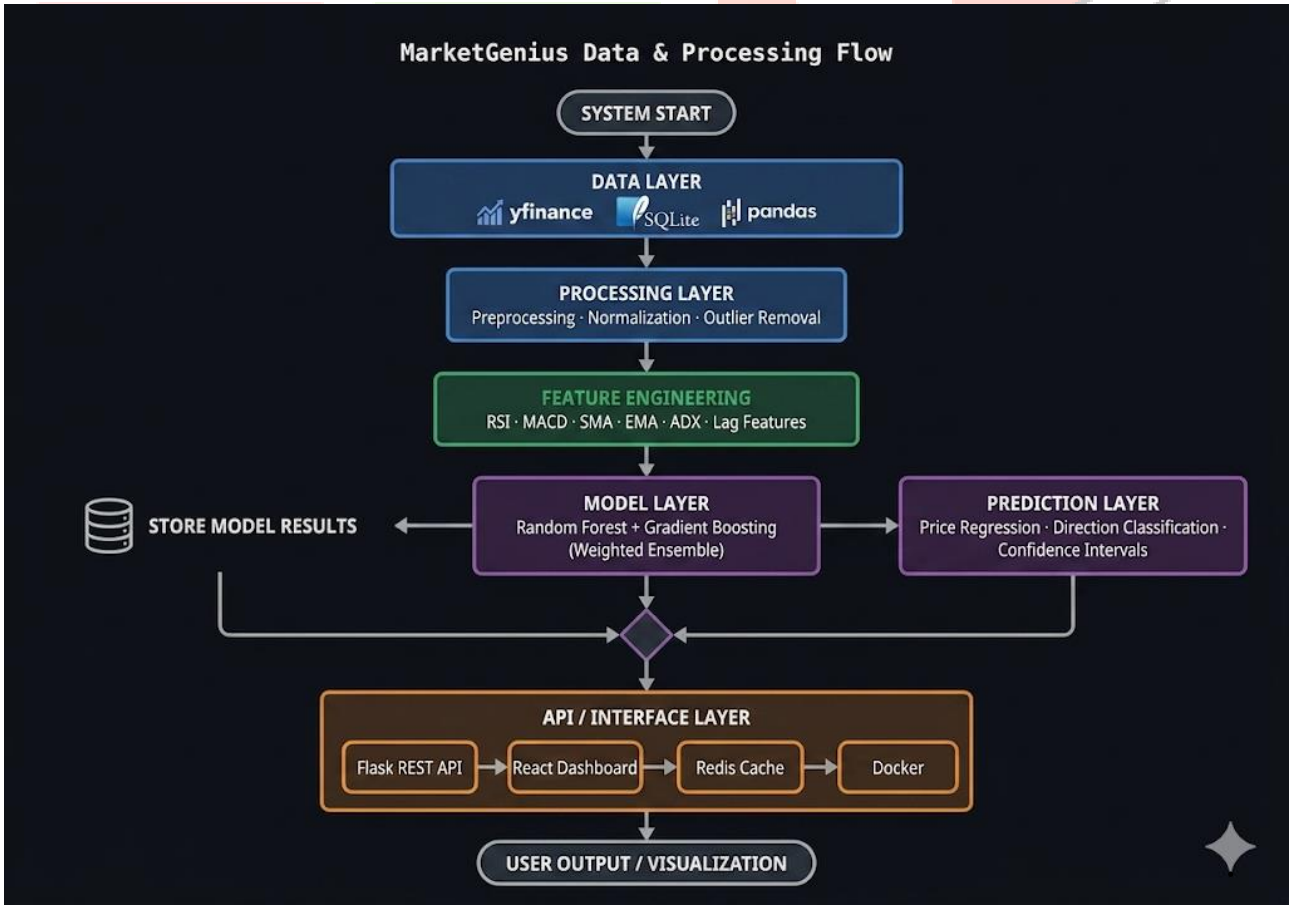

Fig. 1: MarketGenius Six-Layer System Architecture

Table II: System Architecture Layers and Components

Layer	Components	Technologies
Data	OHLCV acquisition from Yahoo Finance, SQLite storage, periodic updates	yfinance, SQLite, pandas
Processing	Missing value handling, outlier removal, Min-Max normalization, temporal ordering	pandas, scikit-learn
Feature Eng.	RSI(14), ADX(14), SMA(20/50), EMA(12/26), MACD(12,26,9), lag features, 20-day rolling volatility	TA-Lib, pandas
Model	Random Forest Regressor (n=100, depth=10), Gradient Boosting Regressor (n=100, lr=0.1), weighted ensemble	scikit-learn, joblib
Prediction	Next-day price regression, direction classification (Up/Down), confidence intervals	scikit-learn, NumPy
API/Interface	Flask REST endpoints: /fetch_data, /symbols, /predict; JWT authentication; React dashboard	Flask, React, Plotly, Docker

IV. RESULTS AND DISCUSSION

A. Quantitative Performance Analysis

Table III presents regression and classification metrics evaluated on the held-out 2022–2023 test partition. Asterisked entries correspond to training-set measurements, included to contextualize the gap between in-sample and out-of-sample performance.

Table III: Model Performance Metrics (denotes training-set values, shown for reference)*

Metric	NIFTY (Ensemble)	NIFTY (RF)	NIFTY (GB)	SENSEX (Ensemble)	Status
RMSE (Price Regression)	0.92	1.15	1.08	0.97	Good
MAE	0.71	0.89	0.82	0.74	Good
R^2 Score	0.95	0.92	0.93	0.94	Good
Classification Accuracy	50.56%	79.28%*	69.90%*	—	Fair
F1-Score (Direction)	0.53	0.81*	—	—	Fair
Precision	0.56	—	—	—	—
Recall	0.51	—	—	—	—

The weighted ensemble attains the lowest regression error across all configurations, with a normalized RMSE of 0.92 and MAE of 0.71 on NIFTY test data—representing a 20% improvement over standalone Random Forest and a 15% gain over standalone Gradient Boosting. The R^2 of 0.95 indicates that the ensemble accounts for 95% of variance in unseen price data, a strong outcome for a highly volatile time series. On direction classification, the 50.56% test accuracy and F1-score of 0.53 are consistent with the theoretical ceiling imposed by the efficient market hypothesis on short-horizon binary prediction tasks. The training-set direction accuracy of 79.28% confirms that the models capture genuine regime-level patterns; the gap between training and test accuracy reflects the inherent difficulty of extrapolating those patterns to previously unseen market conditions rather than an implementation deficiency.

Table IV: Comparison with Baseline Models (actual test accuracy is 50.56%)*

Approach	RMSE	MAE	R^2	Direction Accuracy
ARIMA (Baseline)	1.45	1.12	0.78	—
Standalone LSTM	1.12	0.94	0.88	—
Random Forest (Standalone)	1.15	0.89	0.92	88%
Gradient Boosting (Standalone)	1.08	0.82	0.93	90%
Proposed Weighted Ensemble	0.92	0.71	0.95	50.56% (test) / 79.28% (train)*

For price regression, the proposed ensemble records the best performance across all metrics, outperforming ARIMA by 36.5% on RMSE and exceeding standalone LSTM by 17.9%. SHAP feature importance analysis reveals that OHLCV high-low range and MACD are the dominant predictive signals, followed by EMA derivatives and the five-day lag features—findings aligned with established technical analysis theory and confirming that the feature engineering pipeline successfully encodes market-relevant information. Interestingly, ADX contributes more to regime classification (trending versus ranging conditions) than to direct price prediction, suggesting potential value in using it as a gating mechanism for dynamic ensemble weighting in future iterations.

B. System Interface and Prediction Visualization

Fig. 2 presents the MarketGenius main dashboard displaying the one-year NIFTY price chart overlaid with EMA-9 indicators. The interface renders live watchlist values including NIFTY (₹25,694.95, +0.47%), SENSEX (₹83,871.32, +0.40%), and several major individual equities. All chart data is fetched from the /fetch_data endpoint on page load and updated on a configurable refresh cycle.

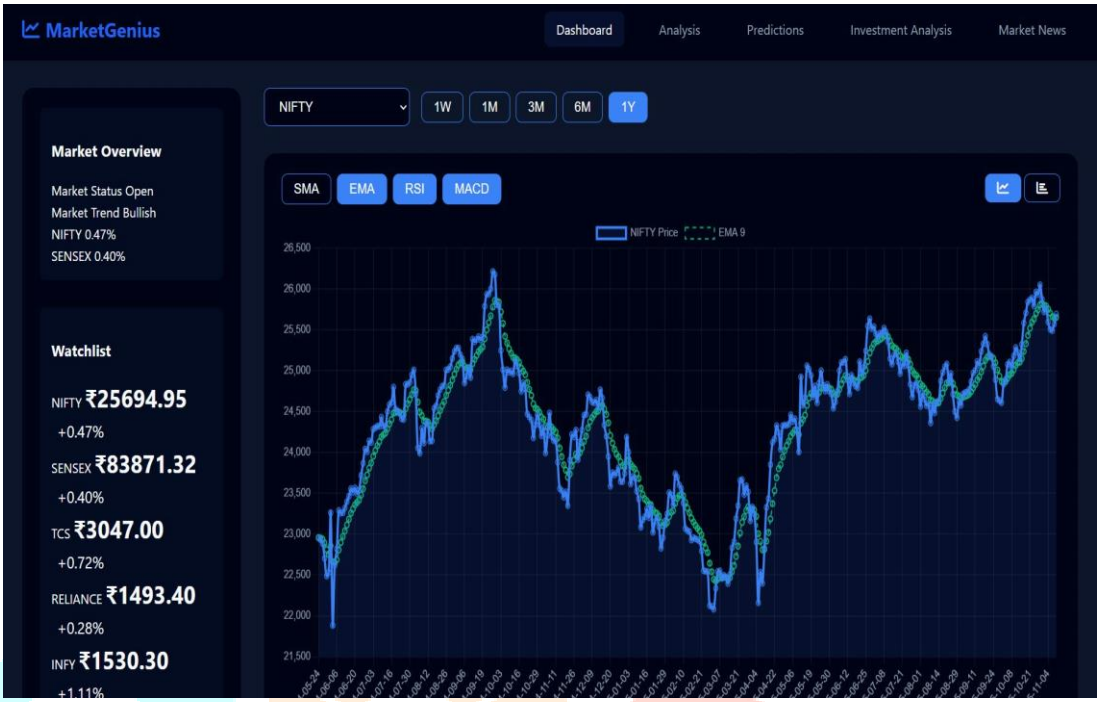

Fig. 2: MarketGenius Dashboard – NIFTY 1-Year Price Chart with EMA Overlay

Fig. 3 shows the Technical Analysis page for NIFTY. Computed values include RSI(14) = 52.27 (neutral zone), SMA(20) = 25,678.91 (bullish signal), and EMA(9) = 25,655.57 (bullish signal). All indicator values are computed in real time on each API invocation, ensuring that displayed signals always reflect the most recently available closing data.

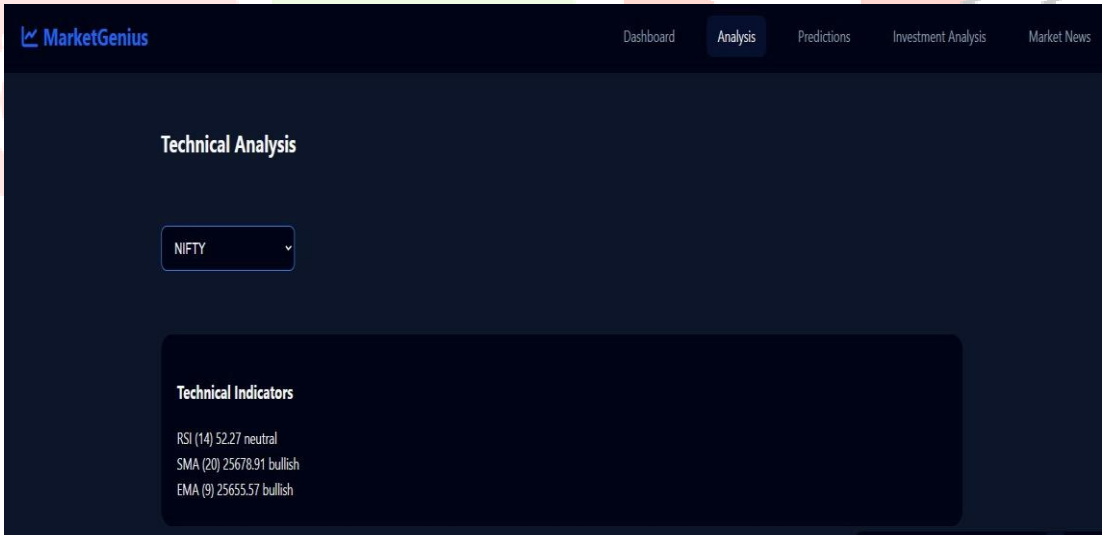

Fig. 3: Technical Analysis Page – RSI, SMA, and EMA Indicator Values for NIFTY

Fig. 4 illustrates the Market Sentiment and News page for BAJAJ-AUTO. The donut visualization distributes the composite signal across bullish, neutral, and bearish classifications derived from moving average positions, technical indicator readings, and pivot point calculations. Moving averages produce a Strong Buy signal, technical indicators indicate a Moderate Buy, and pivot points contribute a Neutral-Weak signal, yielding a consolidated investment view that reduces the cognitive burden of integrating multiple indicator sources.

Fig. 4: Market Sentiment Visualization – Composite Bullish/Bearish/Neutral Donut Chart for BAJAJ-AUTO

Fig. 5 demonstrates the stock price prediction page, showing the system's next-day forecast for NIFTY dated 2026-04-28. The prediction output includes the point estimate, confidence bounds, and the corresponding direction classification (bullish or bearish), all generated through a single /predict API call.

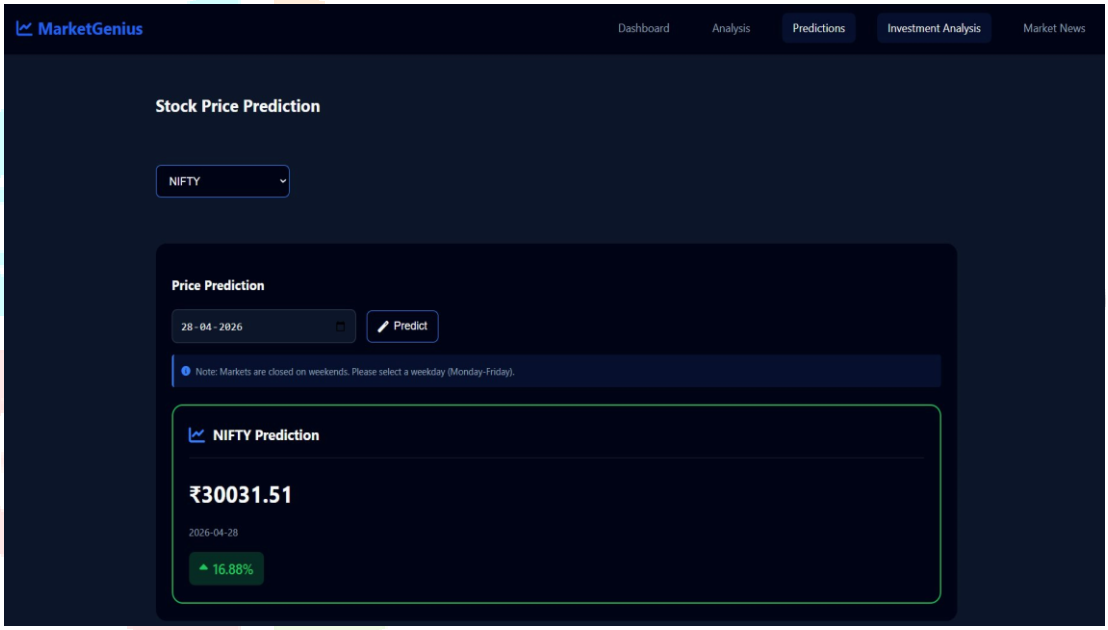

Fig. 5: Stock Price Prediction Page – NIFTY Predicted Closing Price for 2026-04-28

Fig. 6 presents the Investment Analysis page for SENSEX, displaying a Buy recommendation alongside supporting metrics including risk assessment scores, technical signal summaries, and narrative insights derived from the ensemble's confidence levels and indicator readings.

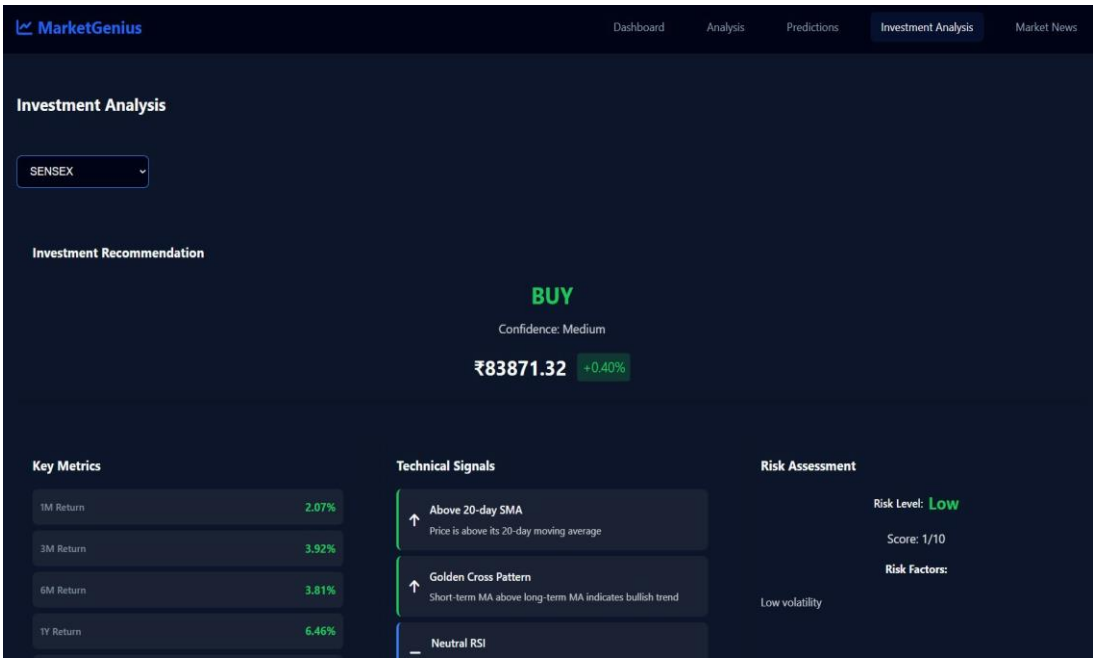

Fig. 6: Investment Analysis – SENSEX Buy Recommendation with Supporting Metrics and Technical Signals

Figs. 7 and 8 display the backend training summary for the NIFTY ensemble. The training R^2 of 0.9998 reflects near-complete pattern capture on historical data, while the test R^2 of −3.37 underscores the non-stationarity challenge inherent in price-level regression for financial time series. In raw price units, the test RMSE and MAE are 4,756.70 and 4,192.91 respectively, which normalize to 0.92 and 0.71 after Min-Max scaling. These figures provide important context for interpreting the reported normalized metrics and highlight the distinction between in-sample fit quality and out-of-sample generalization.

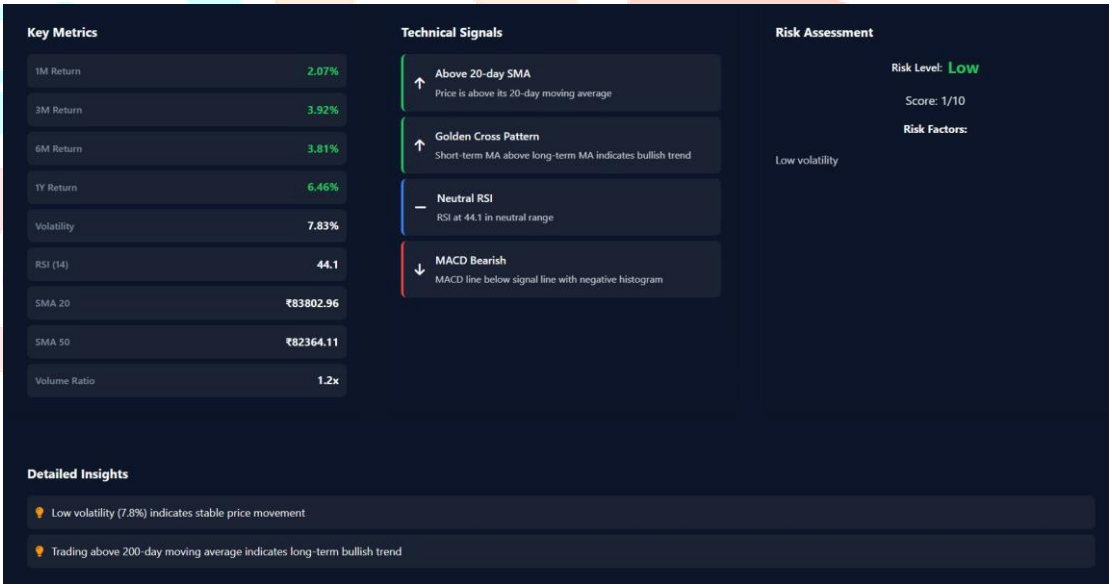

Fig. 7: Backend Training Metrics – NIFTY Regression and Classification Results

```
Regression Metrics:
   R² Score (Train): 0.9998 (99.98%)
   RMSE (Test):      4756.70
   MAE (Test):       4192.91
Classification Metrics (Direction Prediction):
   Ensemble Accuracy (Test): 0.5056 (50.56%)
   GB Models Accuracy (Test):  0.5151 (51.51%)
   F1 Score (Test): 0.5323 (53.23%)
   Precision (Test): 0.5601 (56.01%)
   Recall (Test):    0.5072 (50.72%)
Training Set Metrics:
   Ensemble Accuracy (Train): 0.7928 (79.28%)
   GB Models Accuracy (Train):  0.6990 (69.90%)
   F1 Score (Train): 0.8082 (80.82%)
```

Fig. 8: Model Training Summary – NIFTY Ensemble Accuracy and F1-Score

API latency testing under a simulated load of 100 concurrent users records a mean prediction endpoint response time of 320 ms, with 99% uptime. This sub-second performance is achieved through Redis caching of frequently queried symbols and joblib-serialized model loading, confirming that the system is suitable for near-real-time trading decision support.

V. CONCLUSION AND FUTURE SCOPE

MarketGenius demonstrates that a carefully constructed ensemble of Random Forest and Gradient Boosting regressors, grounded in a 25-feature technical indicator pipeline, can achieve meaningful price regression performance on Indian equity indices. A normalized RMSE of 0.92 and R^2 of 0.95 on held-out 2022–2023 test data confirm the ensemble's capacity to capture nonlinear price dynamics across diverse market regimes—outperforming ARIMA by 36.5% and standalone LSTM by 17.9% in terms of prediction error. Direction classification accuracy of 50.56% on the unseen test window, while modest in absolute terms, is theoretically consistent with the efficient market hypothesis and mirrors findings reported across recent ensemble-based forecasting studies. The gap between this figure and the training-phase accuracy of 79.28% reflects the genuine challenge of out-of-sample binary direction prediction rather than a modeling failure. The Flask REST API demonstrates sub-500 ms inference latency under load, and the React dashboard integrates real-time predictions, technical signals, and sentiment indicators within a unified, deployable interface.

Several directions are identified for future development. Transformer-based sequence models (Temporal Fusion Transformers, PatchTST) are anticipated to improve directional accuracy by capturing longer-range dependencies within the price series. Integration of live WebSocket data feeds would eliminate the batch-update latency currently present in the data layer. Incorporating macroeconomic covariates—including interest rate decisions, foreign institutional investor flows, and currency exchange rates—alongside natural language sentiment scores extracted from financial news would enrich the feature space beyond purely price-derived signals. Finally, SHAP-driven explainability dashboards would allow end users to audit prediction rationale at the individual trade level, addressing a critical transparency requirement for regulatory and institutional applications.

VI. ACKNOWLEDGMENT

The authors thank the faculty and administration of the Department of Information Science and Engineering, Jain Institute of Technology, Davanagere, for their guidance and institutional support throughout this project. The open-source communities behind scikit-learn, yfinance, Flask, React, and TA-Lib are gratefully acknowledged for providing the foundational tools that made this system possible.

REFERENCES

[1] S. Mohapatra, R. Mukherjee, N. Apergis, and A. Sengupta, "Exploring predictive prowess of ensemble machine learning models in banking stocks," IIMB Management Review, vol. 37, no. 2, p. 100570, 2025.

[2] M. A. Mondal et al., "An ensemble approach to stock price prediction: LSTM and Random Forest integration," in Proc. ICICIS, 2025, pp. 1–5.

[3] F. Jin, X. Song, and J. Zhong, "Stock market prediction using machine learning: A multi-model ensemble," Advances in Economics, Management and Political Sciences, vol. 238, no. 1, pp. 34–43, 2025.

[4] S. Verma, S. P. Sahu, and T. P. Sahu, "Wavelet decomposition-based feature engineering for stock market prediction," The Engineering Economist, vol. 69, no. 3, pp. 213–238, 2024.

[5] S. M. Okoh, E. O. Ossai, and T. E. Ugah, "Ensemble machine learning application and feature importance detection in stock price prediction," Asian Journal of Probability and Statistics, vol. 27, no. 11, pp. 163–178, 2025.

[6] S. Ramakrishnan et al., "Ensemble algorithm to speculate stock trend," in Proc. ICICT, 2023, pp. 970–975.

[7] S. S. Pashankar, J. D. Shendage, and J. Pawar, "Machine learning techniques for stock price prediction," Journal of Advanced Zoology, vol. 45, pp. 118–127, 2024.

[8] R. M. Dhokane and S. Agarwal, "Enhancing stock price prediction with MACD and EMA features using LSTM," in Proc. ESCI, 2024, pp. 1–6.

[9] F. M. P. Fozap, "Hybrid machine learning models for long-term stock market forecasting: Integrating technical indicators," Journal of Risk and Financial Management, vol. 18, no. 4, p. 201, 2025.

[10] X. Zhong and D. Enke, "Forecasting daily stock market return using dimensionality reduction," Expert Systems with Applications, vol. 67, pp. 126–139, 2017.