MarketMind: Multimodal and Regime-Conditioned Investment Decision Support Using Sentiment Dynamics and Selective Classification

Vedant Gupta[#1], Thejashwi KA[#2], Sushruth Maiah MR[#3] , Prof. Kavitha P[#4]
Students (Authors 1 to 3) and Professor (Author 4), Department of Information Science and Engineering, CMR Institute of Technology, AECS Layout, Brookefield, Bengaluru, 560037, India
[1]vedantgupta1404@gmail.com, [2]sushruthmaiah151@gmail.com, [3]tanya.thejashwi@gmail.com

***Abstract*—Financial market prediction is difficult because price movements are non-stationary, noisy, regime-dependent, and influenced by information that is not contained in historical prices alone. This paper presents MarketMind, an AI-assisted investment decision-support system that combines historical market information, market-context variables, financial-news sentiment dynamics, regime-aware machine learning, selective prediction, and a full-stack software architecture.**

The system supports Apple (AAPL), Alphabet (GOOGL), Microsoft (MSFT), and Amazon (AMZN). Its software architecture consists of a FastAPI-based machine-learning service, a Spring Boot backend, PostgreSQL persistence, and a web-based frontend. The research prediction pipeline extends a conventional price/technical baseline with market and sector context, volatility information, market regimes, and sentiment-dynamics features. An important provenance distinction is that the historical sentiment series used in offline backtests is synthetically generated from price-return shocks; it is not a point-in-time archive of news articles scored by FinBERT. Real FinBERT inference is used for live news processing in the application. Lagged multimodal representations are constructed using strictly backward-looking information. Binary hurdle targets use a 10-trading-day horizon, while an additional three-class formulation represents BUY, HOLD, and SELL outcomes.

A purged chronological holdout is used for the main research benchmark, with model fitting restricted to the pre-2022 period and out-of-sample evaluation performed over 2022 – 2023. The experiments additionally investigate selective classification, in which predictions are executed only when model conviction exceeds a predefined confidence budget. Under a fixed top-10% conviction policy, the AMZN multimodal ensemble achieved 62.00% directional accuracy across 50 selected observations and an annualized simulated Sharpe ratio of 1.69 after a 5-basis-point round-trip transaction cost. In an AAPL feature ablation, adding the historical sentiment-proxy dynamics increased selected-observation accuracy from 41.89% to 58.11% and the simulated Sharpe ratio from -1.97 to 0.70 under a fixed top-15% policy. Because the historical sentiment series is return-conditioned and synthetic, this result does not establish the predictive value of actual historical news or FinBERT-scored articles.

The results indicate that multimodal and selective approaches can provide useful decision-support signals, but performance is stock-dependent and concentrated at limited coverage. The system is therefore positioned as an experimental investment decision-support platform rather than a guaranteed stock-prediction or financial-advisory system.

Keywords*—*Investment Decision Support, Multimodal Learning, Financial Sentiment, FinBERT, Market Regimes, Selective Classification, Stock Market Prediction, Machine Learning

I. INTRODUCTION

Financial markets are dynamic systems affected by historical price behavior, market-wide movements, sector conditions, volatility, corporate events, and investor reaction to financial information. These characteristics make stock-market prediction substantially different from conventional supervised learning problems. A model may identify useful statistical relationships while still producing unstable performance across stocks and market regimes.

Traditional price-based forecasting systems generally focus on historical numerical observations. Deep-learning approaches such as Long Short-Term Memory (LSTM) networks have been widely investigated for sequential financial data because recurrent architectures can model temporal dependencies. However, price history alone does not explicitly represent broader market context or information contained in financial news.

Financial natural-language processing provides an informative modality alongside quantitative price metrics [1]. Prior investigations have demonstrated that combining financial news sentiment with deep sequential models enhances directional price forecasting over short horizons [1], [2]. Subsequent research has explored large language models for sentiment-driven trading strategies [2], context-enriched movement prediction with dissemination tracking [3], and specialized transformer architectures for financial text classification [4]. Further studies have analyzed domain-specific LLM representations and predictive benchmarks across equity markets [5]. In MarketMind, this NLP pipeline is deployed for live news scoring in production; conversely, the offline experimental benchmark uses a synthetically calibrated return-shock proxy to evaluate feature mechanics without historical textual leakage. These studies motivate the integration of textual information with numerical market features while maintaining clear empirical boundaries between live inference and historical backtesting.

MarketMind was initially developed around a price-based LSTM prediction pipeline and a separate FinBERT sentiment-analysis service. The research version of the system extends this architecture into a multimodal framework. Instead of relying on a single closing-price sequence, the research pipeline combines technical indicators, market and sector context, volatility information, market-regime variables, and historical sentiment-

proxy dynamics. Unlike the live service, the offline historical sentiment series is synthetically generated from price-return shocks rather than from archived news articles. The resulting representation is evaluated with ensemble models and regime-aware mixtures of experts.

An additional objective is to address a practical limitation of financial prediction systems: a model does not necessarily need to issue a trade decision on every observation. Selective classification allows the system to abstain from low-confidence predictions and execute only a predefined fraction of high-conviction signals. This paper therefore evaluates both full-coverage predictive performance and selected-signal performance.

The main contributions of this work are:

- An end-to-end investment decision-support architecture integrating a FastAPI machine-learning service, Spring Boot backend, PostgreSQL database, and web frontend.
- A multimodal market representation combining technical indicators, market and sector context, volatility information, regime variables, and historical sentiment-proxy dynamics, alongside live FinBERT news scoring in the deployed application.
- A regime-aware mixture-of-experts approach for adapting predictions to different market conditions.
- A strict chronological evaluation procedure with purging of training observations whose forward labels cross the temporal training boundary.
- An evaluation of selective classification using fixed conviction budgets and probability-threshold policies.
- An empirical analysis showing that performance is heterogeneous across stocks and that high selective accuracy may occur at substantially reduced coverage.

II. Background and Motivation

A. Financial Time-Series Prediction

Stock prices contain temporal dependencies, volatility clustering, trends, reversals, and responses to market-wide movements. These properties have motivated the application of statistical, machine-learning, and deep-learning methods to financial forecasting.

To capture multi-period dependencies in financial time series, sequential architectures utilize internal gating mechanisms designed to mitigate vanishing gradient effects across temporal intervals [1]. In the baseline MarketMind architecture, an LSTM network processes rolling windows of normalized closing prices to model directional momentum.

However, a model trained only on price history can omit information such as market-wide risk, sector movements, volatility regimes, and financial-news sentiment. The research extension therefore treats price history as one modality within a broader representation

B. Financial-News Sentiment

Financial news provides critical signals regarding corporate earnings, product developments, mergers, regulatory interventions, and executive transitions that reshape market expectations [6], [7]. Domain-specific language models and topic-aware representations transform unstructured financial text into quantitative sentiment signals [7], [8]. Specialized domain adaptations, such as BioFinBERT for clinical and regulatory inflection points [6], demonstrate the utility of targeted textual representations in high-volatility equity environments.

Instead of representing sentiment using only a single article-level score, MarketMind's sentiment feature engine represents multiple temporal statistics, including positive and negative sentiment-like scores, net score, magnitude, moving averages, momentum, acceleration, polarization, and price-sentiment divergence. For the historical offline benchmark, these are proxy features generated by an autoregressive shock process conditioned on short-term returns, not article-level FinBERT outputs from archived historical news.

C. Market Regimes

Market behavior can vary substantially between bullish, bearish, sideways, low-volatility, high-volatility, and crisis conditions. A single global model may not respond equally well across all such conditions.

MarketMind therefore derives market-regime features using broader market information. Trend regimes are based on relationships between the S&P 500 price, moving averages, and recent returns. Volatility regimes are derived from VIX levels. These regime variables are used both as predictive features and, in the regime-aware model, to route observations toward specialized experts.

D. Selective Prediction

Conventional classification evaluates a model over all available test observations. In investment decision support, however, abstention can be meaningful. Whereas standard classification mandates an output for every query instance, selective classification—historically formalized as classification with a reject option—enables a system to abstain whenever posterior uncertainty exceeds an acceptable risk tolerance [9]. In quantitative decision support, abstention represents a disciplined trading policy: capital is deployed exclusively when the model's estimated conviction surpasses a calibrated certainty threshold.

MarketMind evaluates this idea using two approaches. Fixed quantile policies allocate a predefined prediction budget, such as the top 10% or top 15% of model conviction. Parametric confidence sweeps instead vary a probability threshold and measure the resulting accuracy, coverage, and simulated trading performance.

III. Problem Statment

A practical investment decision-support system should process heterogeneous financial information while avoiding future-information leakage during model evaluation. It should also distinguish predictive performance from trading-oriented performance and account for the fact that a model may perform differently across securities and market conditions.

The problem addressed by MarketMind is therefore defined as follows:

> Given historical market observations and the sentiment-proxy information available up to time t, construct a backward-looking multimodal representation and

estimate the direction or decision class of a stock over a future 10-trading-day horizon, while evaluating whether selective execution of high-conviction predictions provides useful decision-support signals under an out-of-sample chronological test.

The system is not designed to guarantee future returns, automate brokerage execution, or provide regulated financial advice.

IV. System Architecture

MarketMind is implemented as a full-stack distributed application with a separate machine-learning service.

A. Application Layers

The architecture contains four principal components:

1. **Web Frontend:** Provides stock visualization, prediction information, sentiment information, decisions, and portfolio-related interfaces.
2. **Spring Boot Backend:** Provides REST APIs, application logic, persistence, prediction orchestration, and the rule-based decision engine.
3. **PostgreSQL Database:** Stores stocks, historical prices, financial news, sentiment results, predictions, decisions, portfolio information, and transactions.
4. **FastAPI Machine-Learning Service:** Provides prediction and financial sentiment functionality and contains the research multimodal prediction pipeline.

The high-level data flow is:

$$\textit{Market Data} + \textit{Financial News} \rightarrow \textit{Feature Processing} \rightarrow \textit{ML Prediction} \rightarrow \textit{Spring Boot Decision Layer} \rightarrow \textit{PostgreSQL} + \textit{Frontend}$$

The separation between the application backend and the Python machine-learning service allows the predictive models to evolve independently of the Java application layer.

B. Supported Securities

The system currently evaluates four technology-sector securities:

TABLE I. SUPPORTED SECURITIES

Symbol	Company	Stock ID
AAPL	Apple Inc.	1
GOOGL	Alphabet Inc.	4
MSFT	Microsoft Corp.	5
AMZN	Amazon.com Inc.	6

Historical market observations cover the 2015 - 2023 period used by the research experiments. The exact number of usable observations varies with feature warm-up periods, missing values, and forward-target construction.

V. Data Acquisition and Feature Engineering

A. Historical Market Data

Historical OHLCV data for model training and the 2022--2023 out-of-sample benchmark is obtained through the yfinance interface and cached as Parquet files. Finnhub is used by the Spring Boot backend for dashboard candle visualization, not as the source of the research training dataset. The research feature pipeline combines historical price and volume information with broader market and sector series.

The broader market context includes:

- S&P 500 index (^GSPC),
- Nasdaq Composite index (^IXIC),
- CBOE Volatility Index (VIX), and
- Technology Select Sector SPDR Fund (XLK).

These series provide information about market direction, market-relative performance, volatility, and sector behavior.

B. Technical Features

The technical feature group includes return, volatility, momentum, trend, and volume variables. Examples include one-, three-, five-, ten-, and twenty-day returns; high-low range; open-close return; rolling volatility; ATR; RSI; MACD components; rate of change; price-to-moving-average ratios; moving-average relationships; volume change; volume ratios; and OBV slope.

These variables are intended to capture multiple aspects of short- and medium-term market behavior rather than relying only on the raw closing price.

C. Market and Sector Context

Market-relative features measure the relationship between the individual stock and broader benchmarks. For a horizon k, the relative return with respect to the S&P 500 is represented as

$$R_{t,k}^{rel} = R_{t,k}^{stock} - R_{t,k}^{SP500}$$

A rolling market beta is estimated over a 60-observation window:

$$\beta_t = \frac{\mathrm{Cov}_{60}(R^{stock}, R^{SP500})}{\mathrm{Var}_{60}(R^{SP500}) + \epsilon}$$

Relative volatility is represented by the ratio of stock volatility to benchmark volatility:

$$V_t^{rel} = \frac{\sigma_{20,t}^{stock}}{\sigma_{20,t}^{SP500} + \epsilon}$$

Similar excess-return features are calculated with respect to the Nasdaq and the XLK sector benchmark.

D. Volatility and VIX Features

The VIX is incorporated as a market-level risk indicator. Features include VIX level, short-term VIX changes, its relationship to its moving average, rolling percentile information, and categorical volatility flags.

The volatility regime thresholds used by the research pipeline are:

- Low volatility: VIX < 16,
- Normal volatility: 16 ≤ VIX < 25,
- High volatility: 25 ≤ VIX < 35,

- Crisis volatility: VIX ≥ 35.

E. Market-Regime Features

The regime detector identifies broad trend conditions using S&P 500 price and moving-average relationships. A bullish regime is associated with the S&P 500 closing price above its 50-day moving average, the 50-day moving average above the 200-day moving average, and a non-negative recent trend condition. A bearish regime uses the corresponding negative relationships, while observations not satisfying either set of conditions are categorized as sideways.

Trend and volatility states are subsequently combined into composite regime identifiers and indicator variables.

The regime representation is used in two ways: as part of the multimodal feature vector and as a routing signal for the regime-aware mixture model.

VI. Financial-News Sentiment Engine

A. FinBERT Processing

For live news processing, MarketMind uses FinBERT to score financial-news text and obtain positive, negative, and neutral probabilities. The historical offline benchmark does not contain a point-in-time archive of news articles for 2015--2021. Instead, its sentiment-like time series is generated by a synthetic autoregressive shock process conditioned on short-term price returns. Consequently, the historical features described below should be interpreted as sentiment proxies, not as historical FinBERT predictions.

The net sentiment score is defined as

$$S_t^{net} = P_{positive,t} - P_{negative,t}$$

Sentiment magnitude is defined as

$$S_t^{mag} = |S_t^{net}|$$

These values preserve more information than a simple categorical label.

B. Sentiment Momentum

The research pipeline aggregates its historical sentiment-proxy series over time. A seven-day proxy momentum measure compares the current seven-day average with the preceding seven-day average:

$$M_{7,t}^{sent} = \mathrm{Avg}(S_{t-6:t}) - \mathrm{Avg}(S_{t-13:t-7})$$

Additional features represent three-, seven-, and fourteen-day momentum, exponential moving averages, sentiment acceleration, rolling z-scores, polarization, positive-to-negative ratios, and volume-weighted sentiment.

C. Price-Sentiment Divergence

The system also measures disagreement between recent price behavior and the sentiment proxy. A divergence measure is constructed by comparing standardized seven-day proxy values with standardized five-day return:

$$D_t = Z\left(S_{7,t}^{sent}\right) - Z\left(R_{5,t}\right)$$

The research pipeline can flag bullish divergence when recent price performance is sufficiently negative while average sentiment remains positive, and bearish divergence when price performance is sufficiently positive while average sentiment remains negative.

D. Sentiment Data Limitation

A critical limitation is the provenance of historical sentiment. The audited implementation generates the 2015 - 2021 historical series through an autoregressive process whose shocks depend on one- and five-day stock returns, with an added stochastic component. Historical headlines were not retrieved and scored with FinBERT for the backtest. Because the proxy is partly conditioned on price momentum, an observed association between proxy features and future returns cannot be attributed to genuine historical news sentiment. The live application separately scores newly ingested Google News RSS articles with FinBERT. These live capabilities and the historical research proxy must not be conflated.

VII. Multimodal Representation

The research pipeline combines the numerical, market-context, regime, and sentiment modalities into a single backward-looking representation.

For each observation at time *t*, the system constructs the feature vector from information available before the prediction point. The representation is then lag-expanded using lags of 1, 2, 3, and 5 observations.

The feature engine defines 83 base features: 27 technical, 25 market/context, 10 regime, and 21 sentiment-proxy features. The audited 292-feature benchmark expands the 73 non-regime base features across four lags (73 × 4); expanding all 83 base features would produce 332 features. Accordingly, the reported 292-dimensional setup excludes the 10 regime columns from the lag-expanded matrix. The lagging procedure is designed to prevent contemporaneous or future observations from entering the model representation.

Conceptually, the multimodal representation can be expressed as

$$X_t = \left[F_{tech,t}, F_{market,t}, F_{regime,t}, F_{sentiment,t}\right]_{lags}$$

The four modalities are:

- **Technical modality:** price, return, volatility, momentum, trend, and volume features.
- **Market modality:** market indices, sector performance, relative returns, beta, and relative volatility.
- **Regime modality:** trend, volatility, and composite market-regime variables.
- **Sentiment modality:** historical sentiment-proxy values and temporal proxy dynamics; live article-level FinBERT probabilities are used separately in production.

Feature selection and standardization are fitted strictly on the training partition in the reported research experiments.

VIII. Prediction Targets

A. Binary Hurdle Target

The primary binary research target identifies whether the stock reaches a positive 10-trading-day return hurdle:

$$y_t = \begin{cases} 1, & R_{t,t+10} \geq 1.5\%, \\ 0, & R_{t,t+10} < 1.5\% \end{cases}$$

This target should not be interpreted as a symmetric up/down target. It specifically evaluates whether the future return crosses the defined positive hurdle.

B. Three-Class Trading Target

An additional decision-oriented target contains BUY, HOLD, and SELL states:

$$y_t = \begin{cases} BUY, & R_{t,t+10} \geq 1.5\%, \\ HOLD, & -1.5\% < R_{t,t+10} < 1.5\%, \\ SELL, & R_{t,t+10} \leq -1.5\%. \end{cases}$$

This formulation uses the same 1.5% hurdle magnitude for BUY and SELL and a neutral interval between the two thresholds. It is a separate decision-oriented experiment and should not be conflated with the binary positive-return hurdle target.

C. Forward-Label Purging

Because the target uses future returns, samples close to the training boundary can otherwise use prices from the test period when their labels are constructed. To prevent this, training observations whose forward-label end date crosses the training boundary are removed. This purged cross-validation approach maintains strict temporal isolation between training and evaluation windows [10].

The training condition is therefore conceptually:

$$date_t < T_{train} \quad \wedge \quad labelEnd_t < T_{train}.$$

This purging procedure prevents boundary leakage caused by forward-looking labels.

IX. Machine-Learning Models

A. Multimodal Fusion Model

The MultimodalFusionModel combines a histogram-based gradient boosting classifier with a regularized logistic-regression model. The ensemble is designed to capture nonlinear relationships while retaining a regularized linear component.

The probability outputs are combined and calibrated using sigmoid probability calibration. Calibration is important because selective classification relies on model confidence rather than only on the predicted class.

B. Regime-Aware Mixture Model

The RegimeAwareMixtureModel extends the multimodal architecture by using specialized experts for different market conditions.

The model contains:

- a bullish-market expert
- a bearish/high-volatility expert
- a global fallback ensemble for transitional conditions.

The regime detector determines the routing context, while the global model provides a fallback when a specialized regime is not sufficiently defined.

This architecture represents the hypothesis that predictive relationships may differ across market states.

C. Research Ensemble Models

Some fixed-budget experiments use a soft-voting ensemble consisting of tree-based models and regularized logistic regression. In the AMZN high-accuracy experiment, the ensemble consists of HistGradientBoosting, ExtraTrees, and regularized LogisticRegression, followed by mutual-information feature selection.

In the AAPL sentiment-proxy ablation, the ensemble consists of HistGradientBoosting, RandomForest, and LogisticRegression, with the feature space reduced using mutual-information feature selection.

These experiments are treated as research configurations rather than as a claim that one universal model is optimal for every stock.

X. Selective Classification

A. Conviction

For a binary classifier with predicted probability P_t, conviction is represented by distance from the neutral probability:

$$C_t = |P_t - 0.5|$$

Higher values indicate stronger model preference for one class.

B. Fixed Quantile Policies

A fixed quantile policy selects a predefined percentage of observations according to conviction. For a top-*K%* policy, only observations above the corresponding conviction percentile are executed.

This approach has an important methodological advantage: the capacity budget is specified independently of test-set labels. For example, a top-10% policy always selects approximately the highest-conviction 10% of observations.

C. Probability Threshold Policies

The regime-aware experiments also evaluate probability-threshold policies. A threshold τ is applied to model confidence, and only observations satisfying the confidence requirement are selected.

Threshold sweeps produce a coverage-performance curve. Because these curves are generated on the out-of-sample period, individual threshold points must be interpreted as conditional operating points rather than as unbiased estimates of a policy chosen before observing test outcomes.

For this reason, fixed quantile policies are emphasized where the objective is to report a predetermined selective policy.

 ISSN: **3107-6513**

XI. Experimental Design

A. Temporal Split

The research benchmark uses a chronological holdout rather than a random train-test split.

TABLE II. Research Evaluation Periods

Period	Role
2015 – 2021	Model training
2022 – 2023	Untouched out-of-sample test

The audited experiments use training observations through December 15, 2021 after forward-label purging. The test period is January 1, 2022 through January 1, 2024, corresponding to the 2022--2023 market period.

The benchmark is therefore more accurately described as a *purged chronological holdout benchmark*. It is not a rolling or expanding walk-forward experiment in which the model is repeatedly retrained during the test period.

B. Leakage Controls

The following controls were applied in the audited experiments:

- Training observations are restricted to the pre-2022 period.
- Forward-label boundaries are purged at the training boundary.
- Feature selection is fitted only on training data.
- StandardScaler parameters are fitted only on training data.
- The 2022 – 2023 observations are retained as the out-of-sample evaluation period.

The forensic audit found no test-label leakage in the reported experiments.

C. Evaluation Metrics

Classification performance is measured using accuracy, balanced accuracy, precision, recall, F1-score, ROC-AUC, and Matthews correlation coefficient (MCC).

Accuracy is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Balanced accuracy is the mean of sensitivity and specificity:

$$BalancedAccuracy = \frac{TPR + TNR}{2}$$

The Matthews correlation coefficient is used as an additional measure because it accounts for all entries in a binary confusion matrix and is informative under class imbalance.

D. Trading-Oriented Evaluation

For selected signals, the research pipeline computes a simulated trade return over the 10-trading-day horizon. A round-trip transaction cost of 5 basis points is deducted from each executed signal.

The annualized Sharpe ratio is calculated as

$$Sharpe = \frac{\overline{R}_{trade}}{\sigma(R_{trade})}\sqrt{\frac{252}{10}}$$

The resulting Sharpe ratio represents a simulated signal-level statistic rather than the performance of a fully implemented portfolio.

The experiments allow overlapping 10-day trades. Consequently, individual trade returns are not independent, and the reported Sharpe ratios should not be interpreted as equivalent to a live, non-overlapping portfolio Sharpe ratio.

XII. Experimental Results

A. Selective Classification Results

The forensic audit verified the provenance of the headline selective results. The results come from different research configurations and should therefore not be combined into a single universal model claim. The repository audit also found that benchmark summary tables and threshold-specific experiment outputs can report different operating points for the same ticker; the table below records the individually audited headline configurations, not a single common threshold sweep.

These selective results should be interpreted alongside the data-provenance limitation described in Section V: historical sentiment-like features are synthetic and return-conditioned, not derived from archived articles scored by FinBERT. In addition, probability-threshold results are operating points from the reported experimental runs; the manuscript does not claim that these thresholds were chosen using an independent validation set.

The AMZN result is generated by a fixed top-10% conviction budget. Fifty observations are selected from the 491-observation test period, producing 10.18% coverage. The selected signals achieve 62.00% accuracy and a simulated annualized Sharpe ratio of 1.69 after transaction costs.

For AAPL, the fixed top-15% sentiment-proxy configuration selects 74 observations, corresponding to 15.07% coverage. The selected observations achieve 58.11% accuracy and a simulated Sharpe ratio of 0.70. This result is relevant to the feature ablation, but it does not isolate the effect of actual historical news because the historical sentiment inputs are synthetic proxy values.

The MSFT regime-aware binary model achieves 68.82% accuracy at 18.94% coverage, but the associated Sharpe ratio is only 0.19. A higher confidence threshold produces 87.50% accuracy and a Sharpe ratio of 1.73, but coverage falls to only 1.63% (8 selected observations). The three-class configuration provides a separate result with a 2.45 Sharpe ratio at 13.24% coverage.

The GOOGL regime-aware binary model reaches 72.22% accuracy at 11.00% coverage, but its Sharpe ratio at that operating point is -0.27. Separately, the three-class configuration produces a 2.94 Sharpe ratio at only 3.26% coverage. These are distinct experiments and should not be interpreted as one model simultaneously achieving both 72.22% accuracy and 2.94 Sharpe.

 ISSN: **3107-6513**

TABLE III. FORENSICALLY VERIFIED SELECTIVE-CLASSIFICATION RESULTS

Stock	Configuration	Policy	Coverage	Trades	Accuracy	Precision	Sharpe
AMZN	Binary Multimodal Ensemble	Top-10% conviction	10.18%	50	62.00%	62.86%	1.69
AAPL	Binary Sentiment-Proxy Ensemble	Top-15% conviction	15.07%	74	58.11%	51.61%	0.70
MSFT	Binary Regime-Aware Mixture	Threshold $\tau = 0.58$	18.94%	93	68.82%	61.11%	0.19
MSFT	Three-Class Decision Model	Threshold $\tau = 0.425$	13.24%	65	44.62%	67.69%*	2.45
GOOGL	Binary Regime-Aware Mixture	Threshold $\tau = 0.60$	11.00%	54	72.22%	64.29%	-0.27
GOOGL	Three-Class Decision Model	Threshold $\tau = 0.425$	3.26%	16	65.50%	81.25%*	2.94

*For the three-class configurations, the reported precision value in the audited trading output represents empirical trade win rate rather than conventional binary classification precision.

B. *AAPL Historical Sentiment-Proxy Feature Ablation*

The feature ablation evaluates the incremental predictive contribution of progressively richer sentiment-proxy representations. Because historical financial news archives were not processed through the FinBERT model for the pre-2022 training partition, the offline series relies strictly on the autoregressive shock formulation described in Section VI. This experiment therefore isolates the structural utility of incorporating dynamic proxy variables into the feature matrix, without asserting empirical claims regarding the predictive efficacy of actual historical financial headlines.

TABLE IV. AAPL HISTORICAL SENTIMENT-PROXY FEATURE ABLATION

Feature Configuration	Accuracy	ROC-AUC	Sharpe
Price/Technical Only	41.89%	0.4677	-1.97
Technical + Market/Macro	51.35%*	0.5282	-0.04
Technical + Basic Sentiment Proxy	48.65%*	0.5006	-0.46
Multimodal + Full Sentiment Proxy	58.11%	0.5490	0.70

*Accuracy values in the intermediate ablation configurations are reported for the selected top-15% observations.

The audited artifact shows an increase in selected-observation accuracy from 41.89% with price/technical features to 58.11% with the full feature configuration, while the simulated Sharpe ratio changes from -1.97 to 0.70 under the top-15% selective framework. However, the historical sentiment component in this experiment is synthetic and return-conditioned, rather than based on archived articles scored by FinBERT. The ablation therefore measures the incremental effect of the implemented proxy feature configuration; it does not isolate the predictive value of real historical news sentiment. This distinction materially limits the interpretation of the ablation.

C. *Full-Coverage Versus Selective Performance*

Selective performance must be interpreted together with coverage. A high accuracy value obtained from a small number of observations does not represent the same operating condition as a model evaluated on the entire test period.

For example, the GOOGL three-class experiment reaches a 2.94 Sharpe ratio using only 16 selected observations out of 491. Similarly, the MSFT 87.50% accuracy operating point contains only 8 observations.

Consequently, MarketMind reports coverage alongside accuracy and simulated financial metrics rather than presenting selective accuracy as a replacement for full-test performance.

D. *Interpretation of Results*

The experiments show three broad observations.

First, predictive performance is heterogeneous across securities. The multimodal representation does not provide a uniform improvement for every stock.

Second, financial-news sentiment appears to have stock-dependent incremental value. The AAPL ablation provides positive evidence under a fixed selective policy, while the broader sentiment experiments do not justify a universal claim that sentiment always improves prediction.

Third, selective classification can substantially alter the apparent performance of a model. High-conviction subsets can exhibit different accuracy and simulated trading characteristics from the full test set. However, reduced coverage and the small number of selected observations must be considered when interpreting these results.

XIII. SYSTEM IMPLEMENTATION

A. *Spring Boot Backend*

The backend is implemented using Java and Spring Boot. It provides REST APIs for stock information, historical prices, financial news, predictions, decisions, portfolio data, and transactions.

The backend coordinates the application workflow:

1. Receive a stock-analysis request.
2. Identify the requested stock.
3. Retrieve relevant market and news information.
4. Communicate with the FastAPI machine-learning service.
5. Persist prediction and sentiment information.
6. Invoke the decision engine.
7. Persist and return the resulting decision.

B. *FastAPI Machine-Learning Service*

The Python machine-learning service provides the numerical prediction and financial sentiment functionality. The service contains the original price-based prediction workflow as well as the research multimodal pipeline.

The research service includes modules for:

 ISSN: **3107-6513**

- market-context construction,
- market-regime detection,
- financial sentiment processing,
- multimodal feature construction,
- target generation,
- ensemble modeling,
- confidence filtering, and
- benchmark evaluation.

This modular design permits individual research components to be evaluated independently.

C. *Decision Engine*

The application decision engine maps analytical outputs into three decision states:

- **BUY**: the combined application signals indicate a positive condition.
- **HOLD**: available signals do not provide a sufficiently positive or negative indication.
- **SELL**: the combined application signals indicate a negative condition.

The decision engine is an application-layer rule system. It should therefore be distinguished from the research classification targets used in the experimental benchmark.

D. *Database Layer*

PostgreSQL is used as the persistent data layer. The database stores:

- stock information,
- historical OHLCV data,
- financial news,
- sentiment results,
- model predictions,
- investment decisions,
- portfolio holdings, and
- transaction records.

Persistence allows the frontend and backend to retain analytical information independently of the Python machine-learning process.

E. *Frontend*

The frontend is implemented using HTML, CSS, and JavaScript. Interactive stock charts are used to display historical price information. The frontend communicates with the Spring Boot REST APIs and displays prediction direction, prediction confidence, sentiment information, decision output, and portfolio-related information.

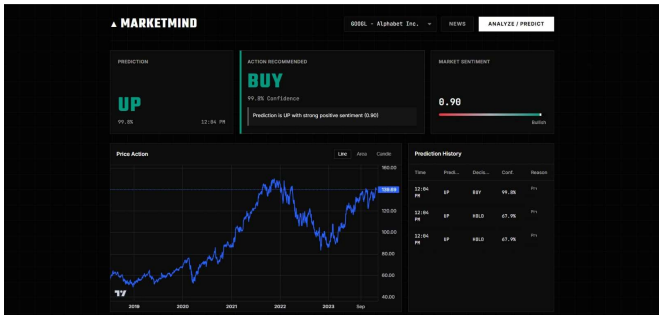

Fig. 1. Marketmind Web Interface

XIV. System Validation

The complete application was tested across its principal integration points. The validation workflow covers communication between the frontend, Spring Boot backend, PostgreSQL database, and FastAPI machine-learning service.

The system supports:

- historical stock-data retrieval,
- stock-price visualization,
- financial-news processing,
- FinBERT sentiment analysis,
- prediction generation,
- investment-decision generation,
- portfolio management, and
- transaction persistence.

The machine-learning research pipeline was additionally evaluated using the chronological experimental protocol described earlier. The forensic audit verified that feature selection and standardization were fitted using training data and that forward-label purging was applied at the training boundary.

XV. Discussion

The results demonstrate why stock-market prediction should be evaluated using multiple dimensions rather than a single accuracy value.

A model may achieve moderate full-test classification performance while producing a smaller subset of predictions with different characteristics. Conversely, a high selective accuracy can be associated with very low coverage. Therefore, accuracy, coverage, precision, ROC-AUC, MCC, and trading-oriented metrics should be considered jointly.

The results also demonstrate that multimodal learning is not automatically beneficial for every security. The addition of market context and synthetic sentiment-proxy features changes the information available to the classifier, but the usefulness of those features depends on the security and period being evaluated.

The AAPL feature ablation indicates that the full historical proxy feature configuration is associated with better selected-subset metrics in that experiment. Because historical sentiment is synthetic and partly driven by price returns, the result cannot be treated as evidence that real news sentiment or FinBERT adds predictive information. A point-in-time historical news corpus and repeated temporal evaluation would be required to test that hypothesis.

The regime-aware results similarly suggest that conditional modeling may alter the behavior of selective signals. However, the high-accuracy operating points for MSFT and GOOGL occur at low coverage, which limits their interpretation as broadly deployable prediction policies.

XVI. Limitations

Several limitations should be considered when interpreting the results.

1. **Limited asset universe:** The experiments focus on four technology-related securities. Results may not

generalize to other sectors, asset classes, or market structures.

2. **Fixed chronological holdout:** The main benchmark is a purged chronological holdout rather than a repeated rolling walk-forward evaluation. Therefore, the reported test period represents one historical regime sequence.
3. **No validation partition in the audited benchmark:** The main benchmark fits the model using the pre-2022 training period and evaluates it on the 2022--2023 holdout. Some threshold-sweep results therefore represent operating points observed on the out-of-sample curve rather than independently selected deployment policies.
4. **Selective coverage:** Some of the strongest numerical results occur at low coverage. For example, the GOOGL three-class Sharpe result is based on 16 observations, while the highest reported MSFT binary accuracy is based on only 8 observations.
5. **Overlapping trades:** The simulated trading evaluation permits overlapping 10-day positions. This creates dependence between observations and can affect the interpretation of annualized Sharpe ratios.
6. **Synthetic historical sentiment:** The 2015 – 2021 sentiment-like series used by the offline benchmark is generated by an autoregressive shock process conditioned on one- and five-day returns. It is not derived from point-in-time archived news headlines scored by FinBERT. Since the proxy incorporates price information, the sentiment ablation cannot establish that actual news sentiment improves prediction and may partly reflect information already present in price features.
7. **Transaction-cost simplification:** The simulation applies a fixed 5-basis-point round-trip cost and does not model slippage, bid-ask spread variation, market impact, taxes, or liquidity constraints.
8. **No live brokerage execution:** The system is a decision-support application and does not execute trades through a brokerage.

XVII. Future Work

The research can be extended in several directions.

- Perform repeated expanding-window and rolling-window walk-forward evaluation.
- Introduce a dedicated validation period for threshold and model-selection decisions.
- Evaluate a larger and more diverse universe of securities.
- Compare additional time-series architectures including GRU, temporal convolutional networks, Transformers, and hybrid models.
- Incorporate richer financial-news event extraction and entity-level representations.
- Evaluate real historical article datasets with explicit publication timestamps and point-in-time availability controls.
- Construct non-overlapping or portfolio-level trading simulations to reduce dependence between individual 10-day signals.
- Report confidence intervals and statistical significance for selective performance.
- Investigate probability calibration and conformal prediction for more reliable abstention decisions.
- Add portfolio-level risk measures including maximum drawdown, volatility, turnover, and exposure constraints.
- Explore explainable AI methods for identifying which market, regime, or sentiment features contribute to individual predictions.
- Investigate deployment scalability and cloud-based infrastructure.

XVIII. Conclusion

MarketMind presents an AI-assisted investment decision-support architecture that combines financial market data, market context, financial-news sentiment, regime information, machine-learning prediction, backend decision logic, persistent storage, and web visualization.

The research extension moves beyond a price-only forecasting approach by constructing a multimodal representation containing technical, market, regime, and historical sentiment-proxy dynamics. The production application separately applies FinBERT to live news; this is distinct from the synthetic historical sentiment series used in the reported offline experiments. A 10-trading-day binary hurdle target and a three-class BUY/HOLD/SELL target are evaluated using a purged chronological holdout covering the 2022 – 2023 period.

The forensic evaluation confirms several useful selective results. The AMZN multimodal ensemble achieves 62.00% directional accuracy at 10.18% coverage with a simulated annualized Sharpe ratio of 1.69 under a fixed top-10% conviction policy. For AAPL, the full-feature sentiment-proxy ablation achieves 58.11% selected-observation accuracy and a simulated Sharpe ratio of 0.70 at 15.07% coverage. Other securities demonstrate that high selective accuracy and high simulated Sharpe can occur under different configurations and coverage levels, emphasizing the need to report the model, target, policy, and coverage together.

The results do not establish a universally superior prediction model or guarantee profitable investment decisions. Instead, they demonstrate a framework for integrating heterogeneous financial information and evaluating whether model confidence can be used to make investment decision support more selective. The system is therefore best understood as an experimental AI-assisted platform that provides analytical signals for human

ISSN: **3107-6513**

interpretation rather than autonomous or professional financial advice.

REFERENCES

[1] S. Halder, "FinBERT-LSTM: Deep learning based stock price prediction using news sentiment analysis," *arXiv preprint arXiv:2211.07392,* Nov. 2022.

[2] K. K. a. G. Germano, "Sentiment trading with large language models," *Finance Research Letters,* vol. 62, p. 105154, Apr. 2024.

[3] Y. L. Y. Liu, N. Wang, H. Yang, B. Zhang and C. D. Wang, "FinGPT: Enhancing sentiment-based stock movement prediction with dissemination-aware and context-enriched LLMs," in *Proc. 1st Workshop Prep. Good Data Generative AI (Good-Data @ AAAI)*, 2025.

[4] Q. C. and H. Kawashima, "Stock price prediction using LLM-based sentiment analysis," in *Proc. IEEE Int. Conf. Big Data (BigData)*, Washington, DC, USA, 2024.

[5] V. Ravi, "Large language models (LLMs) for financial sentiment analysis and market forecasting," in *Proc. Int. Conf. Sustainable Bus. Pract. Inf. Manage. (ICSBPIM)*, 2025.

[6] V. Aparicio, D. G. S. G. Huayamares and Y. Luo, "BioFinBERT: Finetuning large language models (LLMs) to analyze sentiment of press releases and financial text around inflection points of biotech stocks," Jan. 2024.

[7] Z. Wang, Q. Zhang, T. Liu and C. Li, "Analyzing financial news sentiment with NLP to forecast market trends," *International Journal of Engineering and Management Research (IJEMR),* vol. 14, no. 5, pp. 6-11, Oct. 2024.

[8] E. Zhu and J. Yen, "BERTopic-driven stock market predictions: Unraveling sentiment insights," *arXiv preprint arXiv:2404.02053,* Apr. 2024.

[9] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[10] M. L. d. Prado, Advances in Financial Machine Learning, Hoboken, NJ, USA: John Wiley & Sons, 2018.

[11] A. Jadhav and V. Mirza, "Large language models in equity markets: Applications, techniques, and insights," *Frontiers in Artificial Intelligence,* vol. 8, art. 1608365, 2025.