Indian Journal of Commerce, Business & Management (IJCBM)
An International Peer Reviewed, Refereed, Indexed, Quarterly Research journal of Commerce, Business & Management

**ISSN : 3108-057X (Online)**
**3108-1282 (Print)**
**IIFS Impact Factor : 5.625**
Vol.-2; Issue-3 (July-Sept.) 2026
Page No.- 67-83
**DOI : 10.71037/ijcbm.v2i3.06**

https:// ijcbm.gyanvividha.com

**Author's :**

**1. Aanay Shah**
Udgam School For Children,
Ahmedabad-380054, Gujarat, India.

**2. Ajay Goyal**
Institute of Design, Nirma University,
Ahmedabad-382481, India.

Corresponding Author :

**Ajay Goyal**
Institute of Design, Nirma University,
Ahmedabad-382481, India.

# Leakage-Aware Machine Learning for Global Stock Index Direction and High-Volatility Risk Prediction under Walk-Forward and Crisis-Period Validation

**Abstract :** Forecasting short-term equity-market behaviour remains difficult because index movements are noisy, nonlinear, and sensitive to changing market regimes. This study develops a leakage-aware machine-learning framework for classifying next-day direction and high-volatility events across 14 global stock market indices using daily open, high, low, close, and volume data from 2000 to 2026. Features were engineered independently for each index using historical price, volatility, trend, momentum, volume, and calendar information. Future returns and target-derived variables were excluded from the predictor matrix. Next-day direction was modelled as an up/down classification task, while high-volatility events were defined using index-specific 75th percentile thresholds of absolute next-day returns estimated from training-period observations. Candidate models included Logistic Regression, Decision Tree, Random Forest Bagged Trees, Linear SVM, k-Nearest Neighbours, boosting methods, and a SVE. Model selection used validation data, followed by held-out testing, walk-forward validation, index-wise evaluation, and crisis-period robustness assessment. Ensemble-based models showed the strongest performance. For next-day direction prediction, Random Forest Bagged Trees and the SVE achieved accuracy and F1-score of 0.92, with ROC-AUC of 0.97 and PR-AUC of 0.98. For high-volatility prediction, Random Forest Bagged Trees achieved accuracy of 0.95, F1-score of 0.90, ROC-AUC of 0.99, and PR-AUC of 0.98. Walk-forward and crisis-period results indicated that performance remained comparatively stable, although high-volatility prediction was more regime-sensitive than direction prediction. The framework provides a reproducible approach for evaluating market-direction and volatility-risk classification across global indices while avoiding claims about trading profitability or market inefficiency in practical investment settings and applications.

**List of Abbreviations/Symbols**

| Abbreviation/symbols | Full form / Meaning | Abbreviation/symbols | Full form / Meaning |
|---|---|---|---|
| AdaBoost | Adaptive Boosting | TN | True Negative |
| AUC | Area Under the Curve | TP | True Positive |
| DT | Decision Tree | $i$ | Market index identifier |
| EMH | Efficient Market Hypothesis | $t$ | Trading day or time index |
| F1-score | Harmonic mean of Precision and Recall | $j$ | Feature index |
| FN | False Negative | $C_{i,t}$ | Closing price of market index i on trading day t |
| FP | False Positive | $O_{i,t}$ | Opening price of market index i on trading day t |
| GFC | Global Financial Crisis | $H_{i,t}$ | Highest price of market index i on trading day t |
| HVP | High-Volatility Prediction | $L_{i,t}$ | Lowest price of market index i on trading day t |
| kNN | k-Nearest Neighbours | $V_{i,t}$ | Trading volume of market index i on trading day t |
| LR | Logistic Regression | $r_{i,t}$ | Simple daily return of market index i on trading day t |
| LSVM | Linear Support Vector Machine | $l_{i,t}$ | Logarithmic daily return of market index i on trading day t |
| MACD | Moving Average Convergence Divergence | $PV_{i,t}$ | Parkinson volatility estimate for market index i on trading day t |
| NDD | Next-Day Direction | $R_{i,t+1}$ | Next-day return of market index i |
| OHLCV | Open, High, Low, Close, and Volume | $y_{i,t}^{dir}$ | Next-day direction class label for market index i on trading day t |
| PR-AUC | Precision-Recall Area Under the Curve | $y_{i,t}^{vol}$ | High-volatility class label for market index i on trading day t |
| RF-Bag | Random Forest Bagged Trees | $Q_{0.75,i}$ | 75th percentile threshold of absolute next-day returns for market index i |
| ROC | Receiver Operating Characteristic | $x_{i,t,j}$ | Original value of feature j for market index i on trading day t |
| ROC-AUC | Receiver Operating Characteristic AUC | $z_{i,t,j}$ | Standardized value of feature j for market index i on trading day t |
| RSI | Relative Strength Index | $\mu_j^{train}$ | Training-set mean of feature j |

| | | | |
|---|---|---|---|
| SMA | Simple Moving Average | $\sigma_j^{train}$ | Training-set standard deviation of feature j |
| SVE | Soft Voting Ensemble | $K$ | Number of nearest neighbours used in kNN |
| SVM | Support Vector Machine | $\lambda$ | Regularization parameter used in Logistic Regression or Linear SVM |

**1. Introduction :** Forecasting equity-market behavior remains one of the most difficult problems in empirical finance (Fabozzi et al. 2006). Stock indices aggregate the expectations, liquidity conditions, macroeconomic information, risk appetite, and behavioral responses of a large number of market participants (Can Ergün et al. 2022). As a result, index-level price movements are noisy, nonlinear, regime-dependent, and often unstable across time (Alakkari et al. 2026). The classical efficient market hypothesis argues that prices should rapidly incorporate available information, thereby limiting the usefulness of historical price patterns for systematic prediction (Malkiel 2003). However, actual financial markets do not operate under fixed structural conditions (Greenwood and Smith 1997). Market microstructure, monetary policy, global shocks, investor composition, liquidity, and risk preferences evolve continuously. This has motivated a more adaptive view of market efficiency, where predictability may appear, disappear, and reappear across regimes rather than remain constant over the full sample period (Lobão and Costa 2023).

Recent advances in machine learning have renewed interest in return and risk prediction because nonlinear models can capture interactions that are difficult to specify in traditional linear frameworks (Li et al. 2020; Gu et al. 2020). In financial applications, the value of machine learning does not merely lie in replacing econometric models with more complex algorithms. Its real contribution is the ability to examine whether weak, nonlinear, and time-varying patterns exist across large sets of market-derived features. Prior studies have shown that machine-learning methods can improve empirical asset-pricing and return-forecasting tasks when model design, validation, and feature construction are carefully controlled (Li et al. 2020; Gu et al. 2020; Giglio et al. 2022). At the same time, financial machine learning is particularly vulnerable to data leakage, overfitting, unstable parameter selection, and misleading backtest performance (Liu 2024; Bloch 2025; Nikolopoulos 2026). A model that performs well under random splitting may fail when evaluated under a realistic chronological setting, where only past information is available for training and future observations are genuinely unseen.

A major limitation in many market-prediction studies is that they focus narrowly on a single index, a short period, or a single performance metric such as accuracy (Strader et al. 2020; Lee 2025; Nezhad et al. 2026). Such designs are often insufficient for financial inference. Directional accuracy alone does not capture the asymmetry between false positives (FP) and false negatives (FN), nor does it reveal whether a model is useful during market stress. Similarly, return-direction prediction and volatility-risk prediction are economically different tasks (Tehrani et al. 2026; Jain and Kaur 2026). Direction prediction asks whether the market is likely to rise or not, whereas volatility prediction asks whether the next movement is likely to be unusually large, regardless of sign. For traders, portfolio managers, and risk analysts, both questions matter, but they serve different financial decisions. A direction model may support tactical allocation, while a high-volatility model may support exposure reduction, hedging, or risk monitoring.

This study develops a reproducible machine-learning framework for global stock-market index classification using daily open, high, low, close, and volume data across multiple international indices. The modeling pipeline is deliberately structured to reflect finance-specific data constraints, including index-wise feature construction, target generation, leakage-aware predictor filtering, validation-based model selection, and independent holdout evaluation. The primary model-development experiment used a blocked middle-period holdout design. This design was used to evaluate model behavior across temporally separated market blocks without random row-wise splitting. Walk-forward validation was used separately to assess forward-time robustness using expanding historical training windows and later unseen test windows. The evaluation includes accuracy, precision, recall, specificity, Harmonic mean of Precision and Recall (F1-score), balanced accuracy, ROC-AUC, and, for high-volatility prediction, PR-AUC.

In addition to the fixed chronological split, the study implements walk-forward validation using expanding training windows and later test windows. This design better reflects the practical use of forecasting models, where a model is trained on historical data and then evaluated on subsequent unseen periods. Walk-forward testing also allows the study to examine whether model performance is stable across changing market environments or whether predictive ability is concentrated in specific periods. This is especially important in global index forecasting because market behavior changes across crises, monetary cycles, liquidity regimes, and structural shifts in investor behavior.

The final component of the framework evaluates market and model behavior during crisis and non-crisis periods. The analysis explicitly labels major stress windows, including the dot-com crash, the global financial crisis, the COVID-19 market shock, the inflation and interest-rate shock, and a recent market period. For each crisis period and index, the study examines return behavior, annualized volatility, maximum drawdown, and high-volatility event frequency. Model predictions are then merged with crisis labels to assess whether classification performance deteriorates, improves, or changes materially under market stress. This crisis-aware evaluation is essential because a financial model that performs acceptably during calm periods but fails during stress may have limited practical value for risk management.

The contribution of this study is therefore not limited to comparing machine-learning algorithms. It proposes a feature-level leakage-aware, index-level financial machine-learning framework that integrates data auditing, finance-specific feature engineering, index-specific target construction, blocked holdout evaluation, walk-forward validation, multi-metric classification assessment, index-wise reporting, and crisis-period robustness analysis. By jointly modeling next-day direction and high-volatility risk across global indices, the study examines the empirical behavior and limitations of machine-learning-based market prediction. The framework was designed to support reproducible empirical evaluation of market-direction and high-volatility classification across global indices.

**Methodology :** This study developed a time-aware machine learning framework for short-term prediction of global stock market index behavior using daily Open, High, Low, Close, and Volume (OHLCV) data. Each observation represented one market index on one trading day. The market index name was used as the grouping identifier for index-wise processing, feature generation, target construction, and performance evaluation. Any ticker-symbol field, if present, was excluded to avoid duplicate market identifiers. The complete methodological workflow, from raw data preparation to robustness testing, is illustrated in Figure 1.

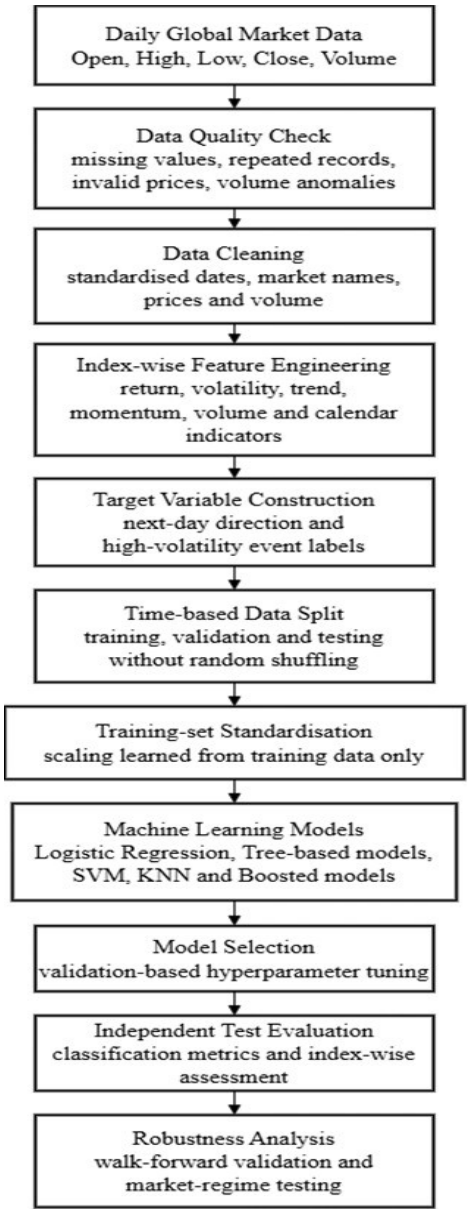


*Figure 1 Overall methodological workflow for time-aware global market prediction.*

The dataset was first screened for structural and financial consistency. This included verification of required fields, missing observations, repeated records for the same market and trading day, non-positive price values, inconsistent high-low price relationships, volume anomalies, and market-wise temporal coverage. Trading dates were converted into chronological date format, price and volume fields were converted into numeric form, and market index names were standardized. Repeated records for the same market on the same trading day were removed by retaining the first valid occurrence. Observations were excluded when the trading date, market identifier, or essential price values were missing, or when open, high, low, or close prices were non-positive. Price consistency was checked by ensuring that the daily high was not lower than the open, low, or close, and that the daily low was not higher than the open or close. Zero-volume observations were retained but flagged because index-level volume is not uniformly reported across all markets and data sources. The dataset structure and data preparation protocol are summarized in **Table 1**.

*Table 1 Dataset description and preparation protocol for the global stock market index dataset.*

| Element | Description | Study value / protocol |
|---|---|---|
| **Dataset structure** | Daily global stock market index panel; each record represents one market index on one trading day. | 14 market indices; 2000-01-03 to 2026-03-25 |
| **Analytical fields** | Trading date, market index name, open price, high price, low price, close price, and volume. | Seven core analytical fields |
| **Volume audit** | Zero-volume observations were retained but flagged because index-level volume reporting varies across sources; negative-volume records were checked. | Zero-volume records: 10,093; negative-volume records: 0 |
| **Dataset size** | Observation count before and after data cleaning. | Raw observations: 91,648; cleaned observations: 91,648; retained fields: 8 |
| **Repeated records** | Repeated records for the same market and trading day were checked before feature construction. | Records flagged/handled: 1 |
| **Missing and invalid price screening** | Records with missing essential information or non-positive open, high, low, or close prices were checked and excluded where applicable. | Records flagged/handled: 1 |
| **Price-consistency audit** | Open, high, low, and close relationships were checked; inconsistent records were flagged for audit. | Flagged records: 1 |
| **Volume audit** | Zero-volume observations were retained but flagged because index-level volume reporting varies across sources; negative-volume records were checked. | Zero-volume records: 10,093; negative-volume records: 0 |

Feature engineering was performed separately for each market index to avoid cross-market information leakage. All predictors were constructed using only current and historical observations. For market index $i$ on trading day $t$, the simple daily return was computed as:

$$r_{i,t} = \frac{C_{i,t} - C_{i,t-1}}{C_{i,t-1}} \quad (1)$$

The logarithmic daily return was computed as:

$$l_{i,t} = \ln\left(\frac{C_{i,t}}{C_{i,t-1}}\right) \quad (2)$$

where $C_{i,t}$ represents the closing price of market index $i$ on trading day $t$. Multi-period returns were calculated over 3-, 5-, 10-, and 20-day windows. Price-structure predictors included intraday high-low range, open-close return, opening gap, candlestick body size, upper shadow, and lower shadow. Rolling volatility was calculated from logarithmic returns over 5-, 10-, 20-, and 60-day windows. Parkinson volatility was estimated as:

$$PV_{i,t} = \sqrt{\frac{\left[\ln\left(\frac{H_{i,t}}{L_{i,t}}\right)\right]^2}{4\ln(2)}} \quad (3)$$

where $H_{i,t}$ and $L_{i,t}$ denote the daily high and low prices, respectively. Intraday range volatility

was also calculated from the high-low spread scaled by the opening price. Trend predictors included 5-, 10-, 20-, 50-, and 200-day Simple Moving Averages (SMAs), close-to-SMA ratios, and SMA-ratio variables. Momentum predictors included Rate of Change (ROC), Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), and short-window momentum variables. Volume predictors included volume change, volume moving averages, relative volume, and price-volume interaction. Calendar predictors included year, month, quarter, and day of week, Monday indicator, and Friday indicator. The first 199 observations of each market index were removed because 200-day SMA-based predictors require a complete historical window. Descriptive checks were performed after cleaning to verify temporal coverage, market-wise availability, and numerical consistency before feature engineering.

Prediction targets were constructed independently for each market index. The next-day return was defined as:

$$R_{i,t+1} = \frac{C_{i,t+1} - C_{i,t}}{C_{i,t}} \tag{4}$$

The next-day direction label was defined as:

$$y_{i,t}^{dir} = \begin{cases} 1, & R_{i,t+1} > 0 \\ 0, & R_{i,t+1} \leq 0 \end{cases} \tag{5}$$

A five-day direction label was also created using the return from trading day $t$ to trading day ($t$+5). High-volatility classification was based on the magnitude of the next-day return. For each market index, the high-volatility threshold was defined as the 75th percentile of absolute next-day returns estimated from the designated training-period observations (Eq. (6)). This index-specific threshold was then used to label unusually large next-day movements.

$$y_{i,t}^{vol} = \begin{cases} 1, & \left|R_{i,t+1}\right| > Q_{0.75,i} \\ 0, & \left|R_{i,t+1}\right| \leq Q_{0.75,i} \end{cases} \tag{6}$$

where $Q_{0.75,i}$ represents the 75th percentile of $\left|R_{i,t+1}\right|$ for market index $i$. Observations for which future target values were unavailable were removed from the corresponding modeling dataset. Future prices were used only for target construction and were excluded from all predictors.

The predictor matrix excluded trading date, market identifier, future returns, target labels, volatility thresholds, and auxiliary splitting fields. Only numeric and logical predictors were retained. A blocked middle-period holdout design was adopted. The validation and test windows were placed in the middle of the sample, while the remaining years before and after these windows were used for model training. Specifically, observations from 2000–2012 and 2017–2026 were used for training, observations from 2013–2014 were used for validation, and observations from 2015–2016 were reserved for final testing. This design was used to evaluate model behavior across temporally separated market blocks while avoiding random mixing of observations.

During validation-based tuning, feature standardization was fitted using the training data and then applied to the validation and test data. After hyper parameter and threshold selection, final models were refitted using the combined training and validation observations where enabled, while the test block remained untouched until final evaluation. Constant or undefined standard deviations were replaced with 1 to avoid numerical instability. For feature (j), the standardized value was calculated as:

$$z_{i,t,j} = \frac{x_{i,t,j} - \mu_j^{train}}{\sigma_j^{train}} \tag{7}$$

where $x_{i,t,j}$ is the original value of feature $j$ for market index $i$ on trading day $t$, $\mu_j^{train}$ is the

training-set mean, and $\sigma_j^{train}$ is the training-set standard deviation. The same training-derived parameters were applied to validation and test data. Constant or undefined standard deviations were replaced with 1 to avoid numerical instability.

For next-day direction prediction, the candidate models included Logistic Regression with ridge regularization, Decision Tree, Random Forest using Bagged Trees, Linear Support Vector Machine (SVM), k-Nearest Neighbours (KNN), Gradient Boosting using LogitBoost, and Boosted Trees using Adaptive Boosting (AdaBoost). Hyperparameters were selected using validation data only. Model selection for direction prediction used the combined validation behavior of F1-score and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). For HVP, Logistic Regression, DT, Random Forest using Bagged Trees, Linear SVM, KNN, and AdaBoost-based Boosted Trees were trained. Because high-volatility observations were comparatively less frequent, model selection used the combined validation behavior of F1-score and PR-AUC. The modeling and validation protocol is summarized in **Table 2**.

*Table 2 Summary of the modeling, validation, and robustness-assessment protocol used for next-day direction and HVP*

| Methodological component | Description | Study-specific setting |
|---|---|---|
| **Prediction tasks** | Supervised classification was used for next-day direction and high-volatility event prediction. Five-day direction was prepared as an auxiliary target. | Main tasks: next-day direction and high-volatility classification |
| **High-volatility threshold** | High-volatility events were defined using a market-specific threshold based on absolute next-day return. | Index-specific 75th percentile of absolute next-day returns estimated from the designated training-period observations. |
| **Predictor matrix** | Trading date, market identifier, future returns, target labels, threshold fields, and split-helper fields were excluded; only numeric and logical predictors were retained. | Leakage-aware predictor filtering |
| **Data partition** | A blocked middle-period holdout design was used for model validation and final testing. | Training: 2000-2012 and 2017-2026; validation: 2013-2014; test: 2015-2016 |
| **Scaling and class weighting** | Z-score standardization was fitted on training data only and applied to validation and test data; class-balanced weights were enabled where supported. | Training-derived scaling; class weighting enabled |
| **Candidate model families** | Logistic Regression, DT, Random Forest using Bagged Trees, Linear SVM, k-Nearest Neighbours, LogitBoost, AdaBoost, and SVE. | Model families compared using validation data |
| **Validation-based selection** | Hyperparameters and classification thresholds were tuned only on validation data. Final models were refitted after model selection. | Composite validation objective; validation-based threshold tuning within 0.20–0.80; final refit after model selection. |

| | | |
|---|---|---|
| **Robustness checks** | Temporal and regime robustness were assessed using expanding-window validation and predefined market-regime windows. | Walk-forward validation and crisis-period assessment |
| **Evaluation metrics** | Performance was assessed using classification metrics defined in the methodology equations. | Accuracy, Precision, Recall, Specificity, F1-score, Balanced Accuracy, ROC-AUC, and PR-AUC |

The final test set was used only after feature preparation, standardization, hyper parameter tuning, and model selection were completed. Performance was quantified using Accuracy, Precision, Recall, Specificity, F1-score, Balanced Accuracy, ROC-AUC, and PR-AUC where applicable. Let True Positive, True Negative, False Positive, and FN be represented by $TP, TN, FP, FN$, respectively. The evaluation metrics were defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

$$Specificity = \frac{TN}{TN + FP} \quad (11)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (12)$$

$$Balanced\ Accuracy = \frac{Recall + Specificity}{2} \quad (13)$$

Temporal robustness was assessed using expanding-window walk-forward validation. The training period was progressively expanded, and each trained model was evaluated on a later unseen period. The validation windows were 2000-2010 to 2011-2012, 2000-2012 to 2013-2014, 2000-2014 to 2015-2016, 2000-2016 to 2017-2018, 2000-2018 to 2019-2020, 2000-2020 to 2021-2022, and 2000-2022 to 2023-2026. In each window, feature standardization was refitted using only the corresponding training period. Walk-forward validation was performed using Logistic Regression, Random Forest using Bagged Trees, Linear SVM, and Boosted Trees.

Market-regime robustness was examined using predefined crisis windows. These included the Dot-com crash, Global Financial Crisis (GFC), Coronavirus Disease 2019 (COVID-19) market shock, inflation and interest-rate shock, and the recent market period. Observations outside these windows were treated as normal-period observations. Where out-of-sample model predictions were available, predictions were aligned with the corresponding market-regime labels using trading day and market index identifier, enabling regime-wise robustness assessment.

After leakage-aware feature filtering, the final analytical sample size and split-wise distribution for each prediction task are reported in Supplementary Table S1. The detailed blocked temporal partitioning used for training, validation-based tuning and held-out testing is provided in Supplementary Table S2. The split-wise class distributions for next-day direction,

five-day direction, and HVP are provided in Supplementary Table S3. Index-wise high-volatility thresholds and event-rate summaries are reported in Supplementary Table S4.

**Results :** Table 3 presents the held-out test performance for next-day direction prediction. RF-Bag and the SVE achieved the strongest overall performance, with accuracy, balanced accuracy, recall, and F1-score of 0.92, supported by high discrimination values for ROC-AUC and PR-AUC. Ensemble Boosted Trees also performed strongly, whereas linear and distance-based classifiers showed comparatively lower, but still consistent, predictive performance.

*Table 3 Held-out test performance of machine-learning models for next-day direction prediction.*

| Model | Accuracy | Balanced Accuracy | Precision | Recall | Specificity | F1 Score | ROC AUC | PR AUC | Threshold |
|---|---|---|---|---|---|---|---|---|---|
| RF-Bag | 0.92 | 0.92 | 0.93 | 0.92 | 0.92 | 0.92 | 0.97 | 0.98 | 0.45 |
| SVE | 0.92 | 0.92 | 0.93 | 0.92 | 0.92 | 0.92 | 0.97 | 0.98 | 0.46 |
| AdaBoost | 0.90 | 0.90 | 0.91 | 0.90 | 0.90 | 0.90 | 0.97 | 0.97 | 0.48 |
| LogitBoost | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 | 0.96 | 0.96 | 0.49 |
| DT | 0.87 | 0.87 | 0.88 | 0.87 | 0.87 | 0.88 | 0.95 | 0.96 | 0.51 |
| LR | 0.87 | 0.87 | 0.88 | 0.87 | 0.87 | 0.87 | 0.95 | 0.96 | 0.52 |
| Linear SVM | 0.86 | 0.86 | 0.87 | 0.86 | 0.86 | 0.87 | 0.95 | 0.95 | 0.54 |
| kNN | 0.86 | 0.86 | 0.87 | 0.86 | 0.86 | 0.86 | 0.94 | 0.95 | 0.55 |

Table 4 presents the held-out test performance for high-volatility event prediction. RF-Bag produced the best overall results, with an accuracy of 0.95, F1-score of 0.90, balanced accuracy of 0.94, ROC-AUC of 0.99, and PR-AUC of 0.98. Boosted trees also showed strong performance, while LR, linear SVM, and kNN achieved lower recall and F1-score, indicating weaker detection of high-volatility events.

*Table 4 Held-out test performance of machine-learning models for high-volatility event prediction.*

| Model | Best Hyperparameters | Accuracy | Precision | Recall | F1 Score | Specificity | Balanced Accuracy | ROC AUC | PR AUC |
|---|---|---|---|---|---|---|---|---|---|
| RF-Bag | NumTrees = 200 | 0.95 | 0.88 | 0.91 | 0.90 | 0.96 | 0.94 | 0.99 | 0.98 |
| AdaBoost | Cycles = 200, LearnRate = 0.100 | 0.94 | 0.85 | 0.90 | 0.87 | 0.95 | 0.93 | 0.99 | 0.98 |
| DT | MaxNumSplits = 100 | 0.93 | 0.82 | 0.87 | 0.84 | 0.95 | 0.91 | 0.99 | 0.97 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| LR | Lambda = 0.0001 | 0.90 | 0.77 | 0.84 | 0.80 | 0.93 | 0.88 | 0.98 | 0.95 |
| Linear SVM | Lambda = 0.1 | 0.90 | 0.75 | 0.82 | 0.78 | 0.92 | 0.87 | 0.98 | 0.95 |
| kNN | K = 3 | 0.88 | 0.71 | 0.80 | 0.75 | 0.90 | 0.85 | 0.97 | 0.93 |

Walk-forward validation results are presented in Table 5. RF-Bag produced the strongest average performance for both prediction tasks. For next-day direction prediction, it achieved a mean F1-score of 0.873 with high ROC-AUC and PR-AUC values of 0.982 and 0.985, respectively. For HVP, it also ranked highest, with a mean F1-score of 0.737, balanced accuracy of 0.887, ROC-AUC of 0.987, and PR-AUC of 0.963. The lower F1-score for HVP reflects the greater difficulty of identifying relatively infrequent risk events compared with directional movements.

*Table 5 Walk-forward validation performance of candidate models for next-day direction and HVP.*

| Task | Model | Mean Accuracy | Mean F1 | Mean Balanced Accuracy | Mean ROC AUC | Mean PR AUC |
|---|---|---|---|---|---|---|
| High Volatility (mean) | RF-Bag | 0.88 | 0.74 | 0.89 | 0.99 | 0.96 |
| | Boosted Trees | 0.86 | 0.71 | 0.87 | 0.98 | 0.95 |
| | LR | 0.84 | 0.67 | 0.85 | 0.98 | 0.93 |
| | Linear SVM | 0.84 | 0.66 | 0.84 | 0.97 | 0.93 |
| Next Day Direction (mean) | RF-Bag | 0.86 | 0.87 | 0.86 | 0.98 | 0.98 |
| | Boosted Trees | 0.85 | 0.86 | 0.85 | 0.98 | 0.98 |
| | LR | 0.82 | 0.83 | 0.82 | 0.97 | 0.97 |
| | Linear SVM | 0.81 | 0.82 | 0.81 | 0.96 | 0.97 |

The best-performing model for each prediction task is summarized in Table 6. RF-Bag was selected for both next-day direction and HVP, indicating stable comparative performance across expanding temporal validation windows. The direction model showed a narrow F1-score range across windows, whereas the high-volatility model showed greater variability, suggesting that risk-event prediction is more sensitive to changing market regimes.

*Table 6 Best-performing model by prediction task based on walk-forward validation.*

| Task | High Volatility (mean) | Next Day Direction (mean) |
|---|---|---|
| Model | RF-Bag | RF-Bag |
| Accuracy | 0.88 | 0.86 |
| F1 | 0.74 | 0.87 |
| Balanced Accuracy | 0.89 | 0.86 |
| ROC AUC | 0.99 | 0.98 |
| PR AUC | 0.96 | 0.98 |

Crisis-period robustness results are presented in Table 7. The selected models maintained strong classification performance during predefined crisis windows, with mean F1-scores of 0.957 for next-day direction prediction and 0.944 for HVP. The high-volatility model achieved particularly high mean recall, indicating strong sensitivity to risk-event detection during stressed market

conditions.

*Table 7 Crisis-period robustness of the selected models for next-day direction and HVP.*

| Prediction Task | Next Day Direction | High Volatility |
|---|---|---|
| Selected Model | DT | kNN |
| Total Observations | 6489 | 6489 |
| Mean F1 Score | 0.96 | 0.94 |
| Mean Recall | 0.95 | 0.97 |
| Mean Precision | 0.96 | 0.92 |
| Mean Balanced Accuracy | 0.96 | 0.97 |
| Mean Specificity | 0.96 | 0.97 |

Figure 2 summarizes the held-out test behavior of the candidate models for next-day direction prediction and HVP. For next-day direction, RF-Bag and SVE showed the strongest overall performance, with high classification scores and clear separation in ROC space. For HVP, RF-Bag again performed best, with strong F1-score, ROC-AUC, and PR-AUC, indicating reliable detection of unusually large next-day movements. The precision-recall behaviour further supports the suitability of ensemble-based models for the more risk-sensitive high-volatility task.

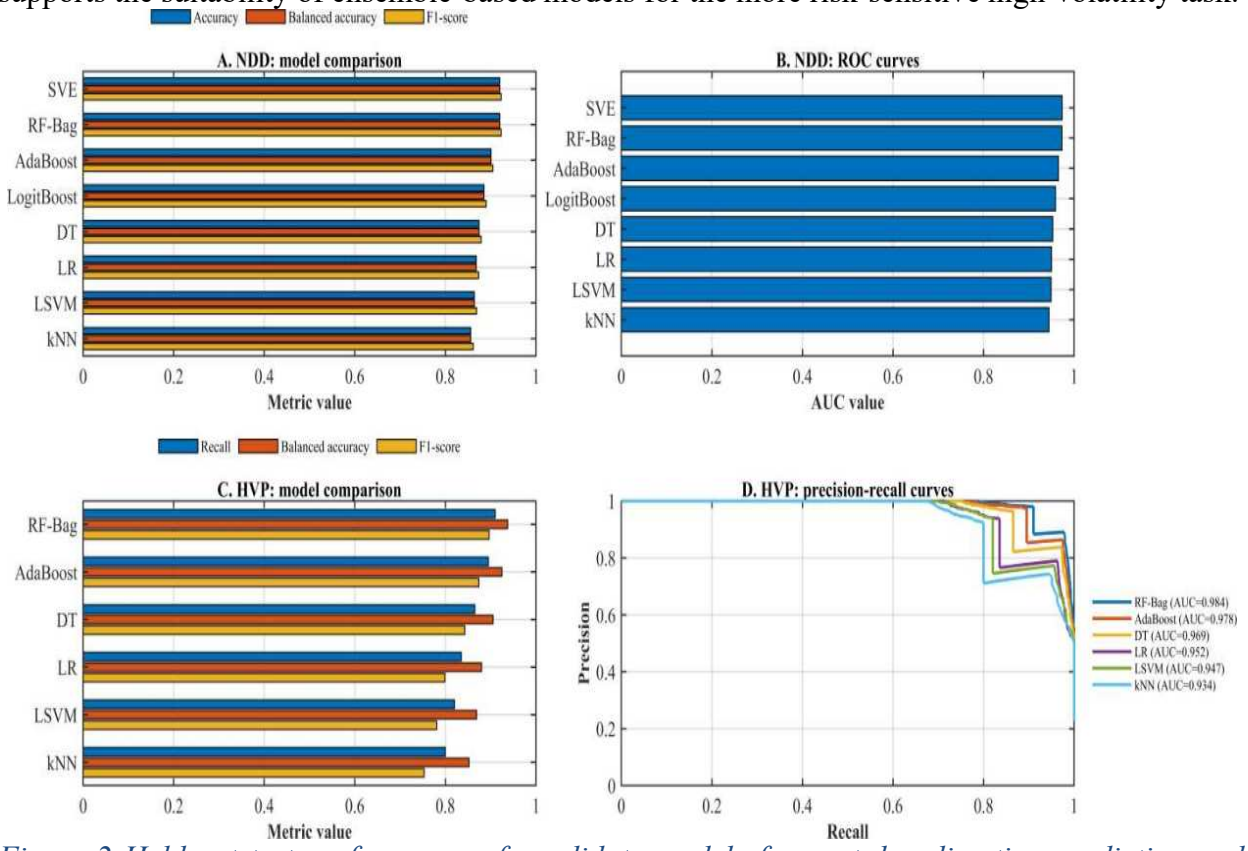


*Figure 2 Held-out test performance of candidate models for next-day direction prediction and HVP.*

Figure 3 shows the index-wise performance of the selected models. The results indicate that predictive performance was not limited to a single market, although some variation across indices was observed. The direction model showed broadly consistent accuracy across markets, while the high-volatility model showed stronger variation in F1-score, recall, and PR-AUC, reflecting the greater difficulty of detecting rare risk events across heterogeneous global indices.

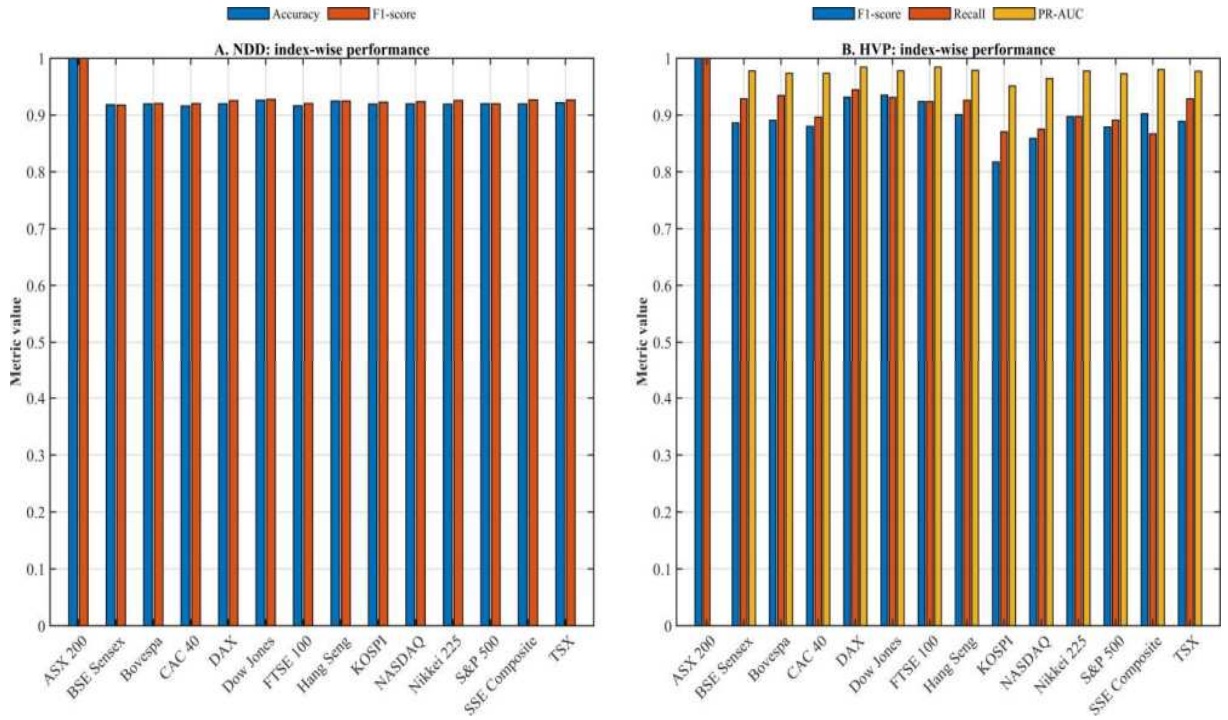


*Figure 3 Index-wise predictive performance of the selected next-day direction and HVP models.*

Figure 4 presents the temporal and crisis-period robustness of the selected models. Walk-forward validation showed that the models maintained stable performance across expanding validation windows, supporting their temporal generalisability. During crisis-period assessment, the selected models retained strong F1-score and recall, particularly for high-volatility detection, suggesting that the framework remained effective under stressed market conditions.

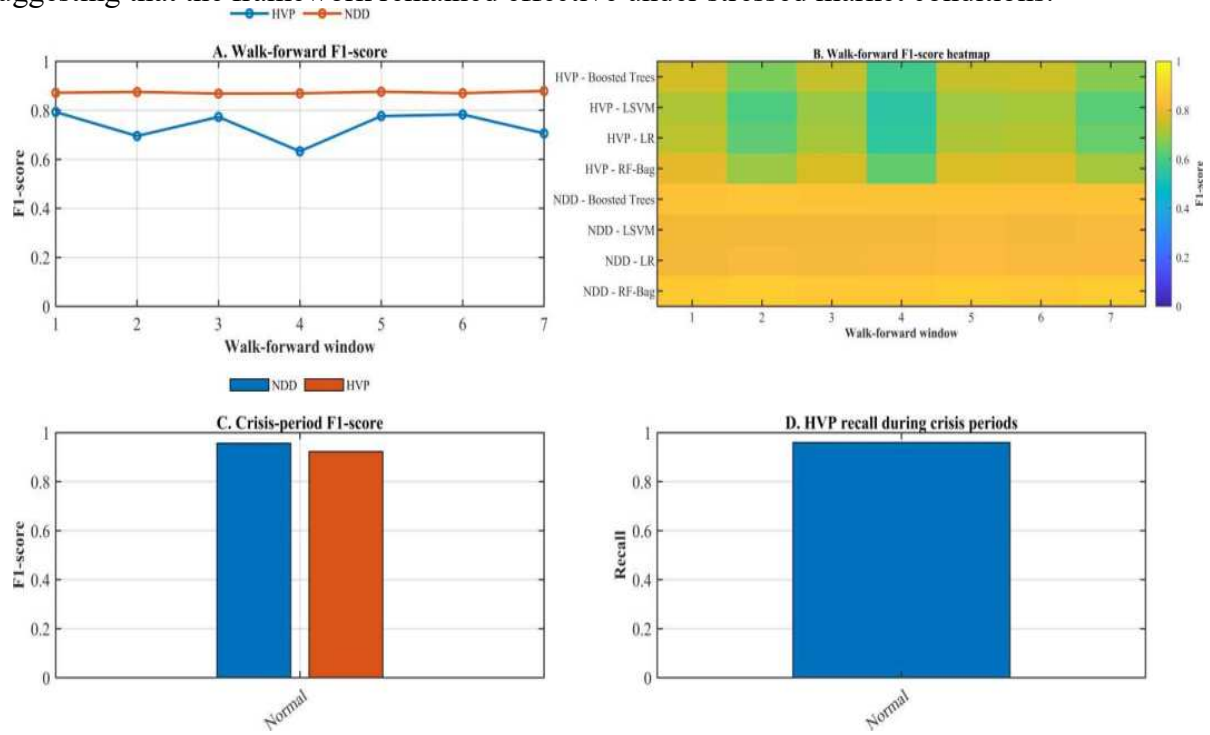


*Figure 4 Temporal and crisis-period robustness of the selected next-day direction and HVP models.*

**Discussion :** This study examined whether machine-learning models trained on historical market-derived features can classify next-day direction and high-volatility events across global stock market indices under a time-aware validation framework. The findings indicate that predictive performance was not uniform across model classes, prediction tasks, or validation settings. Ensemble-based models generally performed better than linear and distance-based classifiers, particularly in the held-out test experiments and walk-forward validation. This pattern is consistent with the expectation that index-level market behavior contains nonlinear interactions among return, volatility, trend, range, momentum, volume, and calendar features that may not be fully captured by simpler parametric classifiers.

For next-day direction prediction, RF-Bag and SVE achieved the highest held-out performance, with balanced classification metrics and strong ROC-AUC and PR-AUC values. The closeness between accuracy, balanced accuracy, recall, specificity, and F1-score suggests that the models did not simply benefit from one dominant class, but maintained comparatively even performance across upward and downward movements. This is important because directional classification in financial markets can easily appear useful when class imbalance or threshold effects are not examined carefully. The performance of AdaBoost and LogitBoost also remained strong, while LR, LSVM, and kNN produced lower but still coherent results. The difference between ensemble and non-ensemble models suggests that the useful predictive information in the engineered features may be distributed across multiple weak signals rather than concentrated in a small number of linearly separable predictors.

The HVP task provides a more risk-oriented perspective. Unlike direction prediction, high-volatility classification does not ask whether the market will rise or fall, but whether the next movement will be unusually large. The results show that RF-Bag performed best for this task, followed by AdaBoost and DT. The higher performance of tree-based methods is meaningful because high-volatility events may be associated with threshold-like behavior, interactions between recent volatility and momentum, and market-specific nonlinear responses. In this setting, recall and PR-AUC are particularly important. A model that misses high-volatility episodes may be less useful for risk monitoring even if its overall accuracy is acceptable. The leading high-volatility models achieved strong recall and PR-AUC, indicating that they were able to identify a substantial proportion of risk-event observations while retaining positive-class discrimination.

The contrast between next-day direction prediction and HVP is one of the central findings of the study. Directional prediction produced stronger F1 stability, whereas HVP showed lower F1 values in walk-forward validation despite high ROC-AUC and PR-AUC. This difference is expected. High-volatility events are less frequent, more regime-dependent, and more sensitive to changing market conditions. The lower F1-score for HVP should therefore not be interpreted as model failure. Rather, it reflects the greater difficulty of event detection when the positive class is more selective and economically distinct. The combination of lower F1 but high ROC-AUC and PR-AUC suggests that the models retained useful ranking ability, even when the final classification threshold faced a more difficult precision-recall trade-off.

The index-wise analysis further strengthens the interpretation of the results. Figure 2 shows that direction-prediction performance was broadly consistent across markets, suggesting that the framework was not driven by a single index. However, high-volatility performance varied more strongly across indices. This is financially plausible because volatility dynamics differ across developed, emerging, commodity-sensitive, technology-heavy, and region-specific markets. Market microstructure, trading calendars, liquidity depth, macroeconomic sensitivity,

and investor composition can all affect how volatility clusters and how quickly information is absorbed into prices. Therefore, the presence of index-wise variation should be treated as an important empirical finding rather than a weakness. It indicates that a global modelling framework can provide useful general structure, but market-specific calibration remains relevant for risk-event prediction.

Walk-forward validation provides a stricter assessment of temporal robustness than a single held-out split. The results show that RF-Bag remained the strongest average model for both prediction tasks across expanding validation windows. This supports the view that the model's performance was not confined to one fixed partition. At the same time, the walk-forward results also show that HVP was less stable than next-day direction prediction. This distinction is important for financial interpretation. A classifier may rank observations well over time, but the economic meaning of the positive class can change as volatility regimes shift. In practice, this implies that high-volatility models may require periodic recalibration, threshold monitoring, or regime-aware updating even when the underlying model family remains suitable.

The crisis-period assessment adds a further layer of evidence. Financial prediction models often perform differently during market stress because correlations rise, liquidity conditions change, volatility clusters, and behavioural responses become more extreme. The reported crisis-period results suggest that the selected models retained strong F1-score and recall during predefined stress windows. The high recall for HVP is especially relevant because risk-monitoring systems are primarily concerned with detecting stress episodes rather than merely maximizing average classification accuracy. However, this result should be interpreted carefully. Strong crisis-period classification does not necessarily imply that the model can forecast all future crises or anticipate structural breaks. It indicates that, within the predefined crisis windows and available out-of-sample predictions, the model retained useful classification behavior.

The findings also have implications for the efficient-market and adaptive-market perspectives discussed in the introduction. The results do not contradict market efficiency in a strong form, nor do they prove persistent exploitable abnormal returns. The models classify short-term direction and volatility states using historical market-derived information; they do not directly demonstrate profitability after transaction costs, slippage, portfolio constraints, taxes, or execution delays. A more appropriate interpretation is that the engineered features contain limited but detectable statistical information about short-term market states, and that this information may vary across time and regimes. This aligns more closely with an adaptive-market view, where predictability is conditional, time-varying, and dependent on market environment rather than constant.

From a methodological standpoint, the study's main contribution lies in combining leakage-aware feature construction, index-specific target labeling, validation-based tuning, held-out testing, walk-forward validation, and crisis-period robustness analysis. This is important because many financial machine-learning studies can overstate performance when random splits, future information leakage, or incomplete validation designs are used. The blocked holdout design provides one controlled test of unseen observations, while walk-forward validation gives a more realistic forward-time assessment. The use of multiple metrics also avoids reliance on accuracy alone, which can be misleading in imbalanced or risk-event settings.

The study should nevertheless be interpreted within clear limits. First, the dataset is based on daily OHLCV information. It does not include macroeconomic releases, interest rates, sentiment, futures-implied information, options-based volatility measures, fund flows, or cross-asset predictors. These additional data sources may improve risk-event detection but would also

introduce new timing and leakage challenges. Second, the analysis evaluates classification performance rather than trading profitability. A model with strong classification metrics may not automatically translate into an economically profitable strategy once costs, liquidity, position sizing, and drawdown constraints are considered. Third, high-volatility labels are based on index-specific historical thresholds. This is appropriate for cross-index comparability, but the threshold definition may influence class frequency and model behavior. Future work could compare fixed, rolling, and regime-dependent volatility thresholds.

Fourth, although crisis-period analysis is valuable, crisis labels are necessarily retrospective. Future stress episodes may differ from past crises in speed, origin, policy response, and market transmission. Therefore, crisis robustness should be viewed as evidence of historical stress-period stability, not as a guarantee of future crisis performance. Finally, the current framework focuses on prediction at the next-day horizon. Longer horizons, multi-class market-state definitions, probabilistic calibration, and portfolio-level decision rules could provide additional insight into practical financial usefulness.

Overall, the results suggest that a carefully designed machine-learning framework can extract useful short-term classification signals from global stock index data, particularly when ensemble models are combined with leakage-aware validation and multi-metric evaluation. The evidence is strongest for the comparative advantage of tree-based ensemble methods and for the importance of evaluating both direction and volatility-risk tasks separately. The study therefore contributes not by claiming universal market predictability, but by demonstrating a reproducible and time-aware framework for assessing when and how machine-learning classifiers retain useful performance across markets, validation windows, and stressed regimes.

**Conclusion :** This study developed a leakage-aware machine-learning framework for short-term classification of global stock market index behavior using daily OHLCV data. By separating next-day direction prediction from high-volatility event prediction, the study addressed two distinct financial forecasting problems: directional movement and risk-event detection. The results showed that ensemble-based models, particularly RF-Bag, produced the most consistent performance across held-out testing, walk-forward validation, and crisis-period assessment. HVP was more challenging and more sensitive to market regimes than direction prediction, but the leading models retained useful recall and discrimination ability. The findings suggest that historical market-derived features contain measurable short-term classification information when evaluated through careful feature construction, validation-based tuning, and time-aware testing. However, the results should be interpreted as evidence of statistical classification performance rather than direct trading profitability. Future work should examine economic value after transaction costs, probabilistic calibration, rolling thresholds, and integration of macroeconomic or cross-asset predictors.

**References :**

1. Alakkari K, Abotaleb M, El-kenawy E-SM, Mishra P (2026) Advanced Deep Statistical Learning Approach for Forecasting Global Economic Policy Uncertainty and Volatility. Comput Econ. https://doi.org/10.1007/s10614-026-11350-7
2. Bloch DA (2025) False Findings in Finance: The Hidden Costs of Misleading Results in the Age of AI
3. Can Ergün Z, Cagli EC, Durukan Salı MB (2022) The interconnectedness across risk appetite of distinct investor types in Borsa Istanbul. Stud Econ Finance 40:425–444. https://doi.org/10.1108/SEF-09-2022-0460

4. Fabozzi FJ, Focardi SM, Kolm PN (2006) Financial Modeling of the Equity Market: From CAPM to Cointegration. John Wiley & Sons
5. Giglio S, Kelly B, Xiu D (2022) Factor Models, Machine Learning, and Asset Pricing. Annu Rev Financ Econ 14:337–368. https://doi.org/10.1146/annurev-financial-101521-104735
6. Greenwood J, Smith BD (1997) Financial markets in development, and the development of financial markets. J Econ Dyn Control 21:145–181. https://doi.org/10.1016/0165-1889(95)00928-0
7. Gu S, Kelly B, Xiu D (2020) Empirical Asset Pricing via Machine Learning. Rev Financ Stud 33:2223–2273. https://doi.org/10.1093/rfs/hhaa009
8. Jain S, Kaur K (2026) AI in Stock Market Forecasting: A Systematic Review of Regional Performance, Crisis Robustness, and Transfer Learning
9. Lee MC (2025) Temporal Fusion Transformer-Based Trading Strategy for Multi-Crypto Assets Using On-Chain and Technical Indicators. Systems 13:474. https://doi.org/10.3390/systems13060474
10. Li Y, Turkington D, Yazdani A (2020) Beyond the Black Box: An Intuitive Approach to Investment Prediction with Machine Learning. J Financ Data Sci 2:61–75. https://doi.org/10.3905/jfds.2019.1.023
11. Liu P (2024) Rethinking Model Risk in Finance
12. Lobão J, Costa AC (2023) The Adaptive Dynamics of the Halloween Effect: Evidence from a 120-Year Sample from a Small European Market. Int J Financ Stud 11:13. https://doi.org/10.3390/ijfs11010013
13. Malkiel BG (2003) The Efficient Market Hypothesis and Its Critics. J Econ Perspect 17:59–82. https://doi.org/10.1257/089533003321164958
14. Nezhad MM, Rouhani S, Mohammadi N, Shahedi A (2026) A unified GRU model for cryptocurrency price prediction and harsh price movement detection using enhanced sentiment analysis. Sci Rep 16:16064. https://doi.org/10.1038/s41598-026-46271-w
15. Nikolopoulos SD (2026) Spurious Predictability in Financial Machine Learning
16. Strader T, Rozycki J, ROOT T, Huang Y-H (2020) Machine Learning Stock Market Prediction Studies: Review and Research Directions. J Int Technol Inf Manag 28:63–83. https://doi.org/10.58729/1941-6679.1435
17. Tehrani MHE, Borujeni AS, Hashemigohar M, et al (2026) A Hybrid Approach Based on a Memory-Instance–Based Gated Transformer (MIGT) Algorithm and Metaheuristic Optimization for Portfolio Management with the Aim of Return Optimization and Risk Control. J Resour Manag Decis Eng 1–20

•