



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85068>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Explainable and Feature-Engineered Machine Learning for Stock Market Forecasting: A Survey of Lexicon-Based, Statistical, and Sentiment-Driven Approaches

Sidharth S¹, Harichandana H S², Albin Tom Thomas³, Akhilesh P A⁴, Mrs. Sagara M R⁵

Department of Computer Science and Engineering (Data Science), St. Thomas Institute for Science and Technology, Kerala, India

Abstract: Forecasting the direction and magnitude of stock price movements remains one of the most persistently difficult problems in applied machine learning, owing to the noisy, non-stationary, and weak-form-efficient character of financial time series. This survey examines three recent studies that approach the problem through complementary lenses: an explainable Decision Tree that builds dynamic, industry-specific lexicons from financial news to classify next-day price-change magnitude [1]; a feature-engineering and portfolio-construction pipeline that derives new price-based predictors via Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and a proposed Statistical Average Method (SAM) before applying eight regression architectures, including Bi-LSTM [2]; and an LSTM-based trend classifier that fuses FinBERT-scored Twitter sentiment with price data to forecast North American and European bank-stock movements during the macroeconomically turbulent year of 2022 [3]. Beyond summarising each study's objectives, methodology, data, and results, this survey situates the three works within the broader and rapidly evolving literature on deep sequence models (LSTM, Bi-LSTM, GRU, and transformer architectures), sentiment analysis with pretrained financial language models such as FinBERT, and explainable artificial intelligence (XAI) in finance. A structured comparative analysis contrasts the three studies along ten dimensions, including data sources, feature engineering, model architecture, explainability, validation design, and reported performance, while explicitly cautioning against directly ranking studies whose prediction tasks and evaluation metrics are not commensurable. A cross-domain case study of an attention-based legislative-vote predictor is retained to examine whether a learned similarity mechanism could improve on the fixed similarity rules used throughout the finance literature reviewed here. The survey's contribution lies in bringing explainable lexicon-based methods, statistically engineered price features, and sentiment-driven deep learning together within a single comparative framework, and in using that synthesis to identify concrete, evidence-grounded gaps around auxiliary-signal fusion, the explainability of high-accuracy sequence models, and cross-market generalisation.

Keywords: stock market forecasting; explainable artificial intelligence; feature engineering; LSTM; Bi-LSTM; transformer; financial sentiment analysis; FinBERT; portfolio optimization; attention mechanism.

I. INTRODUCTION

The task of anticipating how financial markets will move next has long drawn the attention of both practitioners and academics, given its bearing on investment strategy, risk control, and economic policymaking. According to the Efficient Market Hypothesis, current prices already incorporate all publicly available information, so that systematic forecasting should, in principle, offer no advantage; the closely related random walk view likewise treats price changes as essentially unpredictable from historical prices alone [1], [3]. Even so, a considerable body of behavioural and empirical finance research points to information asymmetries, sentiment-driven trading, and market microstructure effects that leave room for at least partially exploitable regularities — a possibility that continues to drive increasingly elaborate applications of machine learning (ML) and deep learning (DL) to financial forecasting [1]–[3].

Three recent studies tackle this problem from markedly different angles. Carta et al. [1] build dynamically updated, sector-specific lexicons out of large volumes of financial news and turn the resulting lexicon-overlap statistics into engineered features for an explainable Decision Tree, deliberately valuing transparency over raw accuracy.

Sami et al. [2] take a different route, augmenting the standard price attributes — open, high, low, close, and adjusted close — with statistically derived variables obtained through PCA, LDA, and a newly proposed Statistical Average Method (SAM), and feeding the enlarged feature set into a wide array of regressors before using the resulting price forecasts to assemble an optimised portfolio. Bacco et al. [3] instead examine how LSTM networks perform when supplemented with sentiment signals mined from the tweets of well-known financial commentators via FinBERT, focusing on bank-stock trend prediction for North American and European institutions during a year marked by unusual macroeconomic and geopolitical strain.

Although the three studies target different outputs — categorical magnitude classification, price-level regression, and binary trend classification, respectively — they converge on a common goal: improving the accuracy, and where feasible the interpretability, of stock-related predictions by supplementing raw price series with additional engineered or externally sourced signals. This shared aim makes the three works natural candidates for a joint review, and this survey brings them together to draw out common threads, contrasting design decisions, and gaps left unaddressed.

Beyond these three focal papers, the wider deep-learning literature on stock forecasting has continued to expand. Recent surveys catalogue several dozen distinct neural architectures applied to the problem, ranging from convolutional and recurrent networks to transformer-based sequence models that use self-attention rather than recurrence to model temporal dependence [6], [11]. Situating the three focal papers within that broader landscape — rather than reading them in isolation — is one of the aims of the expanded literature review in Section 3.

To check whether the broader idea of enriching a base model with textual or relational similarity holds up outside the financial domain, a brief cross-domain case study is also included on a recent legislative-prediction model that applies an attention mechanism to weight similar legislative bills as auxiliary evidence [4]. Although this model operates in an entirely different setting — predicting how individual lawmakers will vote — its central methodological claim, that a learned similarity measure can outperform fixed formulas such as cosine similarity, bears directly on the lexicon-matching and feature-engineering strategies used in the three finance papers reviewed here. It is treated as a pointer toward future work rather than as a fourth subject of the survey in its own right.

The rest of the paper proceeds as follows. Section 2 sketches the background of machine learning applied to stock forecasting. Section 3 reviews related work, including literature published since the three focal studies. Section 4 analyses each focal paper in turn. Section 5 presents a structured summary comparison, and Section 6 extends this into a dimension-by-dimension comparative table. Sections 7 through 9 compare methodology, datasets and experimental design, and reported performance, respectively. Section 10 weighs advantages against limitations. Section 11 presents the cross-domain case study on attention-based similarity. Section 12 states the survey's contribution and scope. Sections 13 and 14 set out research gaps and future directions, and Section 15 closes the paper.

II. BACKGROUND: MACHINE LEARNING FOR STOCK MARKET FORECASTING

Much of the research on stock prediction begins from the friction between the Efficient Market Hypothesis and its behavioural or adaptive counterparts. Carta et al. [1] appeal instead to the Adaptive Market Hypothesis, under which excess returns arise from uneven access to information, so that traders able to gather and interpret varied sources — news, social media, financial ratios — gain a real edge. Bacco et al. [3] make a related point: even though the random walk theory casts doubt on the value of extrapolating from price trends alone, stock movement prediction remains an active research area precisely because richer information sources may still hold predictive value.

On the methodological side, the field has moved from technical-indicator-driven statistical models, through bag-of-words and lexicon-based text representations, toward deep architectures able to learn temporal structure directly from sequential data. Recurrent networks — LSTM and its bidirectional counterpart, Bi-LSTM — became the default building block for this purpose, since their memory cells and gating mechanisms let them retain information over long sequences, a property put to direct use by both Sami et al. [2] and Bacco et al. [3]. More recently, transformer architectures, which model temporal dependence through self-attention rather than recurrence, have been proposed as an alternative capable of capturing longer-range dependencies and of parallelising more efficiently during training [6], [9], though at the cost of typically requiring larger datasets and offering less inherent sequential inductive bias than recurrent models.

At the same time, folding in textual information — whether through hand-built financial lexicons [1] or transformer-derived sentiment scores from models like FinBERT [3], [7] — reflects a growing sense that price data alone may miss economically relevant signals.

FinBERT itself was developed specifically because general-purpose sentiment classifiers, trained on everyday language, misclassify much financial terminology whose polarity is domain-specific (for instance, a rising cost figure is negative in general usage but can be neutral or even positive in a financial-growth context) [7]. Finally, the explicit push for interpretability seen in Carta et al.'s Decision Tree approach [1] mirrors the broader rise of Explainable Artificial Intelligence (XAI), a concern of particular weight in finance given the regulatory and practical need to justify automated trading or advisory decisions [5].

III. LITERATURE REVIEW AND RELATED WORK

Each of the three papers draws on a partly overlapping body of earlier work. Carta et al. [1] trace how NLP-driven financial forecasting has evolved from early bag-of-words and statistical measures toward deep-learning methods, positioning their lexicon-based Decision Tree as an extension of their own earlier rule-based lexicon-matching system, while treating explainability as an explicit design goal consistent with the wider XAI literature.

Sami et al. [2] survey a substantial body of work on feature selection and dimensionality reduction for stock prediction — including chi-square feature testing, PCA-based reduction, and LIME/SHAP-based interpretability — and note that, while much prior work has examined which existing price features carry the most predictive value, comparatively little had, at the time, looked at newly constructed additive or averaged price attributes as predictive features in their own right. This gap is what motivates their SAM/average-based price feature.

Bacco et al. [3] chart the field's progression from purely price-based technical analysis, through neural approaches (including CNNs and LSTM variants paired with econometric models), to the incorporation of textual sentiment drawn from financial news and social media. They single out FinBERT as a transformer model built specifically for financial sentiment analysis, and observe that although several earlier studies link social-media sentiment to price movement, few had systematically examined how such signals interact with different regional markets or with varying amounts of aggregated training data.

Beyond these study-specific antecedents, four broader strands of more recent literature bear directly on this survey and were not available to, or not discussed by, the three focal papers at the time of their writing.

A. Deep sequence and transformer architectures for stock forecasting

Dedicated surveys of deep learning for stock prediction position LSTM and, increasingly, transformer-based architectures as the dominant sequence models in the field. Zou et al. [6] propose a taxonomy spanning stock movement classification, price regression, portfolio management, and event-driven prediction, and catalogue the datasets and evaluation protocols used across this literature, while Jiang [11] provides a complementary survey of data sources, network structures, and implementation details for deep-learning-based prediction more broadly. These surveys corroborate the reliance on LSTM and Bi-LSTM seen in Sami et al. [2] and Bacco et al. [3], while also flagging transformer architectures — which none of the three focal papers uses — as an increasingly common alternative for modelling long-range temporal dependence.

B. Sentiment analysis and hybrid textual–numerical fusion

Financial sentiment analysis has moved from lexicon- and rule-based scoring toward pretrained language models. FinBERT [7], the model Bacco et al. [3] apply to their Twitter corpus, is a domain-adapted version of BERT [12], the bidirectional transformer encoder whose masked-language-model pretraining underlies most subsequent transformer-based text representations used in finance, including the BERT-derived sentiment embeddings in Sonkiya et al. [10] discussed below. FinBERT was developed specifically to address the mismatch between general-purpose sentiment classifiers such as base BERT [12] and the specialised vocabulary of financial text. More recent work has fused transformer-derived textual representations with price-based deep learning more tightly than simple feature concatenation: Sonkiya et al. [10] combine BERT-derived sentiment embeddings with a generative adversarial network for price forecasting, and Lu et al. [8] combine convolutional feature extraction, Bi-LSTM, and an attention mechanism (CNN-BiLSTM-AM) that learns to weight the relative importance of different time steps, reporting improved accuracy over plain LSTM and Bi-LSTM baselines on Shanghai Composite Index constituents. Most recently, hybrid architectures have begun combining recurrent layers, transformer-style attention, federated learning, and sentiment inputs within a single pipeline [9]. Read alongside Bacco et al.'s [3] simpler concatenation of FinBERT scores with price features, this body of work suggests that how sentiment is fused with price information — and not merely whether it is included — may materially affect predictive performance, a question the three focal papers do not examine directly.

C. Explainable AI in finance

The explicit push for interpretability in Carta et al.'s Decision Tree [1] sits within a wider and increasingly systematic literature on XAI in finance. Cernevičienė and Kabašinskas' systematic review of 138 studies published between 2005 and 2022 finds that stock-price prediction is among the most common applications of XAI in finance, alongside credit scoring and fraud detection, and that feature-attribution methods (chiefly SHAP) and rule-based simplification methods such as decision trees are the two dominant explanation strategies [5] — situating Carta et al.'s [1] rule-based Decision Tree within an established methodological family rather than as an isolated design choice. The same review finds that the higher-accuracy deep architectures typical of stock-price forecasting (LSTM, CNN, and their hybrids) are explained, when at all, almost exclusively through post-hoc methods such as SHAP or LIME applied after training, rather than through architectures that are interpretable by design [5] — a gap that applies directly to Sami et al.'s [2] Bi-LSTM and Bacco et al.'s [3] LSTM models, neither of which incorporates any explanation mechanism.

D. Feature engineering and dimensionality reduction

The standard price attributes and technical indicators underlying Sami et al.'s [2] feature-engineering pipeline have long been supplemented by dimensionality-reduction and feature-construction techniques elsewhere in the forecasting literature, including chi-square-based feature screening and PCA-based reduction. Sami et al.'s [2] own characterisation of this prior work — that comparatively little study had examined newly constructed additive or averaged price features relative to feature-selection methods applied to existing indicators — remains, to the extent verifiable from the literature identified for this survey, an accurate description of a genuine and still under-explored gap; a fully independent verification of this claim would require a dedicated systematic search of the feature-engineering literature, which is beyond the scope of the present survey and is flagged accordingly in Section 13.

Read together, these four strands reinforce rather than overturn the original positioning of the three focal papers: price data alone is generally judged insufficient, but the field has not converged on which auxiliary signal, or which fusion mechanism, is most effective — a divergence that motivates the comparative analysis in Sections 5 and 6.

IV. DETAILED ANALYSIS OF SELECTED STUDIES

A. Explainable Lexicon-Based Forecasting (Carta et al. [1])

Carta et al. [1] set out to predict whether the size of a firm's next-day stock price change will be classified as “high” or “low” relative to a fixed cut-off, deliberately choosing magnitude classification over exact price-level forecasting because it is more directly actionable for volatility-aware trading and risk-management decisions. Their pipeline runs through four stages — lexicon construction, feature extraction, prediction, and explanation. Sector-specific lexicons are built dynamically from the Dow Jones “Data, News and Analytics” (DNA) corpus by weighting each term in a news article according to the absolute price move recorded on the following trading day, and averaging these weights over every occurrence of the term. Ten real-valued features are then computed for each company-day pair by measuring the overlap between the lexicon and various time-windowed groupings of news — company-specific versus sector-wide; previous day, week, or month; and, where relevant, historically tied to high versus low variation. A Decision Tree trained on these features is then read directly to generate rule-based explanations, supplemented by illustrative sentences drawn from the news items responsible for triggering each rule.

The choice of a Decision Tree over a higher-capacity model is a deliberate trade-off: the authors are explicit that raw predictive accuracy is secondary to producing an explanation a human analyst or regulator can audit directly, a design stance consistent with the broader finding that rule-based and feature-attribution methods dominate XAI applications in finance [5]. The resulting Decision Tree is shown to match or outperform Random Forest, Gradient Boosting, and Multilayer Perceptron baselines, as well as a simpler unsupervised lexicon-matching method from the authors' earlier work. The authors argue that a Balanced Accuracy of 0.5–0.6 already constitutes a meaningful result, since the broader volatility- and magnitude-forecasting literature generally finds this task considerably harder than typical classification benchmarks. Feature-importance analysis also shows that company-level monthly features carry more predictive weight than sector-level ones, and that high-variation features consistently outweigh their low-variation counterparts — an observation the authors suggest could justify building a dedicated “high-impact” lexicon in future work.

B. Statistical/PCA/LDA Feature Engineering and Portfolio Construction (Sami et al. [2])

Sami et al. [2] approach the problem from the angle of feature engineering combined with portfolio optimisation. Starting from the five standard price attributes — Open, High, Low, Close, and Adjusted Close — the authors construct three additional candidate features: a one-dimensional price compiled via PCA, a one-dimensional price compiled via LDA, and a Statistical Average Method (SAM) price obtained by simply averaging the five conventional attributes.

The underlying rationale is that PCA and LDA condense correlated price attributes into components that capture the dominant modes of joint variation, while the deliberately simple SAM feature tests whether a naive averaging heuristic can rival more sophisticated dimensionality reduction once fed into a sufficiently expressive downstream model. Each candidate is tested for independence using a chi-square test against a 10% significance threshold; the PCA-derived feature narrowly misses this bar (9% independence) and the LDA-derived feature misses it by a wider margin, whereas the SAM/average feature clears it comfortably (24.4% independence). Only the conventional and SAM-augmented feature sets — labelled CFD and AFD — are therefore carried forward into modelling.

Ten stocks — six chosen for high or low beta (volatility) and four for beta near one — are selected through financial-ratio screening (positive EPS, current ratio between one and two) and used to train eight regression models: Support Vector Regression, Decision Tree Regression, Logistic Regression, Random Forest Regression, XGBoost, LSTM, Bi-LSTM, and GRU, each tuned with the Optuna hyper-parameter framework. The March–June 2020 COVID-19 volatility window is deliberately chosen as the test period to put the models under stress, since a feature engineered on calmer historical data may fail to generalise to a genuine tail event. The Bi-LSTM trained on the AFD (SAM-augmented) feature set delivers the best reported accuracy (98.14%) and adjusted R^2 (above 99%), and an overfitting check on a held-out validation window shows validation and test accuracy staying consistent. Feeding these forecasts into the PyPortfolioOpt library, subject to a minimum Sharpe ratio of one and a target annual return above a savings-account benchmark, produces an AFD-based portfolio with a two-month ROI of 55.589% (against 46.76% for the conventional feature set) and a Sharpe ratio of 77.62.

C. LSTM and Sentiment Analysis Under Macroeconomic Uncertainty (Bacco et al. [3])

Bacco et al. [3] cast stock prediction as a binary trend-classification task — positive versus negative next-day price movement — for 26 major bank stocks (17 tied to the US Federal Reserve, 9 to the European Central Bank) across the unusually volatile year of 2022.

Their core model is an LSTM network with stacked recurrent layers, fully connected layers, and a softmax output. Around this base architecture, the authors run a series of controlled experiments to isolate five factors: market region (US versus EU); local training (a separate model per bank) versus global training (data pooled across banks within a region); the span of training data (a full year versus the most recent half-year); the length of the input sequence (5, 10, or 20 trading days); and whether or not FinBERT-derived sentiment features — built from roughly 14,000 tweets by 34 prominent financial commentators and executives, weighted by retweets and likes — are included.

A Pearson correlation analysis shows US banks moving considerably more in tandem (average 0.778, peaking at 0.982 between Goldman Sachs and Morgan Stanley) than EU banks do (average 0.596), which the authors attribute to more unified federal economic policy in the US relative to the greater fiscal and monetary variation across EU member states. This structural gap carries through to modelling choices: global, aggregated training tends to beat local, per-bank training, particularly for EU banks, since pooling data helps average out country-specific policy noise, whereas a few unusually large or small US banks do better with local training. Weighted F1-scores average 0.68 for US banks (reaching roughly 0.80 in the strongest cases) but only 0.447 for EU banks (dropping as low as 0.181 for BNP Paribas). Shorter input sequences (5–10 days) and more recent, half-year training data generally outperform longer sequences and full-year training, though not consistently across every bank. Adding Twitter sentiment brings mixed results: it helps several of the largest US banks by market capitalisation but not uniformly, and offers markedly less benefit for European banks — a pattern the authors trace to the US-centric composition of the sentiment corpus and the way sentiment scores are aggregated.

Considered against the more tightly integrated sentiment-fusion architectures surveyed in Section 3 — for instance the attention-weighted combination used in Lu et al.'s CNN-BiLSTM-AM [8] — Bacco et al.'s [3] concatenation of a single aggregated sentiment score with price features is comparatively simple, and the uneven benefit of sentiment they report may partly reflect this simpler fusion mechanism rather than a fundamental limit on the value of social-media sentiment as a signal.

V. COMPARATIVE ANALYSIS

Table 1 lays out the objective, methodology, dataset, algorithms, evaluation metrics, key findings, and limitations of the three focal papers side by side.

TABLE 1. COMPARATIVE SUMMARY OF THE THREE SURVEYED STOCK-FORECASTING PAPERS.

Paper (Year)	Objective	Methodology	Dataset	Algorithms / Evaluation / Key Results	Limitations
Carta et al. [1] (2021)	Classify whether the next-day price change of S&P 500 firms will be large or small in magnitude, using news-derived lexicons.	Dynamic construction of sector-specific lexicons from news text; ten engineered company/sector features across day, week and month windows and high/low variation groups; Decision Tree classifier with rule-based and sentence-level explanations.	Dow Jones DNA news corpus (2005–2018) with S&P 500 prices across three sectors (IT, Financial, Industrials); ten-fold walk-forward validation.	Models: Decision Tree vs. Random Forest, Gradient Boosting, MLP, unsupervised lexicon baseline. Metrics: Balanced Accuracy; Precision/Recall/F1 for the ‘high’ class. Results: Decision Tree matches or beats every baseline; Balanced Accuracy 0.5–0.6 treated as meaningful; best F1 for the ‘high’ class.	Bag-of-words lexicon misses semantics/context; no explicit event extraction; validated on one national market only; no word-frequency analysis.
Sami et al. [2] (2025)	Raise stock-price forecast accuracy via engineered price features (PCA, LDA, statistical averaging), then build an optimised, lower-risk portfolio.	Three new price features — PCA-Price, LDA-Price, SAM/Average-Price; chi-square independence testing; eight ML/DL regressors tuned via Optuna; portfolio optimisation with PyPortfolioOpt.	Ten stocks (six high/low-beta, four near beta 1) from a pool of 42; investing.com / Investopedia / macrotrends; trained Jan 2010–Mar 2020, tested on COVID-19 window (Mar–Jun 2020).	Models: SVR, Decision Tree Reg., Logistic Reg., Random Forest Reg., XGBoost, LSTM, Bi-LSTM, GRU. Metrics: Accuracy, adjusted R ² , ROI, Sharpe ratio. Results: Bi-LSTM + AFD (SAM) reaches 98.14% accuracy, R ² > 99%; portfolio ROI 55.589% vs. 46.76% (conventional); PCA/LDA features fail chi-square test.	Ignores investor behaviour, news, and macro/commodity variables; favours accuracy over compute speed; no short-horizon regression setup; 2-month portfolio window.
Bacco et al. [3] (2024)	Determine whether LSTM plus Twitter sentiment (FinBERT) improves bank-stock trend prediction in 2022’s high macroeconomic uncertainty, across US and EU markets.	LSTM (stacked recurrent + fully connected + softmax) tested across five factors: region, local vs. global training, training-data span, input sequence length, and presence/absence of FinBERT sentiment.	Yahoo Finance OHLCV for 26 banks (17 US Fed-tied, 9 EU ECB-tied), Dec 2021–Dec 2022; ~14,000 tweets from 34 influential accounts, scored with FinBERT.	Model: LSTM. Metrics: Pearson correlation (inter-bank); weighted F1. Results: US banks more correlated (avg. 0.778) and score higher (avg. F1 0.68, up to 0.80) than EU banks (avg. corr. 0.596, avg. F1 0.447, as low as 0.181); global training helps fragmented EU market; shorter sequences (5–10 days) and recent data perform best; sentiment helps some large US banks but not consistently, less so in EU.	Restricted to bank stocks; Twitter sample drawn from a curated influencer set (sampling bias risk); binary trend classification rather than exact price estimation.

VI. DIMENSION-BY-DIMENSION COMPARISON

Table 1 summarises each paper as a whole; Table 2 complements it by cutting across the three studies along ten specific dimensions of comparison — prediction task, auxiliary signal, feature validation, architecture, explainability, validation strategy, metrics, computational complexity, market coverage, and generalisability — to make the similarities and differences easier to read at a glance.

TABLE 2. DIMENSION-BY-DIMENSION COMPARISON OF THE THREE SURVEYED PAPERS.

Dimension	Carta et al. [1]	Sami et al. [2]	Bacco et al. [3]
Prediction task	Magnitude classification (high/low next-day move)	Price-level regression, then portfolio construction	Binary trend classification (up/down)
Auxiliary signal	Financial-news lexicon overlap	Statistically engineered price features (PCA/LDA/SAM)	FinBERT-scored Twitter sentiment
Feature validation	Feature-importance ranking within the Decision Tree	Chi-square independence test (10% threshold)	Ablation across five experimental factors
Core architecture	Decision Tree (shallow, interpretable)	Eight regressors; best result from Bi-LSTM	LSTM (stacked recurrent + dense + softmax)
Explainability mechanism	Native rule extraction + illustrative news sentences	None reported	None reported
Validation strategy	Ten-fold walk-forward split	Train 2010–2020; test on COVID-19 window; separate over-fit check	Multiple train/test splits per experimental factor
Evaluation metrics	Balanced Accuracy; Precision/Recall/F1	Accuracy; adjusted R ² ; ROI; Sharpe ratio	Pearson correlation; weighted F1
Computational complexity	Low (shallow tree, lexicon lookup)	High (eight tuned models incl. three recurrent variants)	Moderate–high (LSTM across 26 banks × 5 factors)
Market / asset coverage	S&P 500; three US sectors	Ten US-listed stocks across beta profiles	26 banks, US and EU
Generalisability	Untested outside the S&P 500	Untested outside the ten selected stocks / COVID window	Explicitly compared across US vs. EU; not tested beyond banking

Two patterns stand out from Table 2. First, explainability and predictive sophistication trade off sharply across the three studies: the only paper with a native explanation mechanism [1] also uses the shallowest model, while the two papers reporting the highest raw accuracy or F1 [2], [3] provide no explanation mechanism at all. Second, each paper validates its auxiliary signal through a different lens — feature importance, statistical independence testing, or factorial ablation — and none of the three cross-validates its own auxiliary signal using another paper's validation method, which limits how confidently their respective auxiliary-signal claims can be compared.

VII. METHODOLOGIES AND TECHNIQUES

The three papers embody three distinct philosophies for strengthening a base predictive model with auxiliary information. Carta et al. [1] rely on a dynamically updated lexicon built straight from financial news text, matched against documents through simple percentage-overlap statistics — a method that stays transparent and easy to interpret but, as the authors themselves note, cannot capture word order, syntax, or deeper semantic relationships, the familiar limitation of bag-of-words representations. Sami et al. [2] instead derive new numerical price attributes through classical statistical and dimensionality-reduction techniques (PCA, LDA, and plain averaging), checking their value through chi-square independence testing rather than drawing on text at all. Bacco et al. [3] follow a third route, pulling dense sentiment representations out of unstructured social-media text with a pretrained transformer (FinBERT) — capturing far more semantic nuance than lexicon matching, but giving up the direct interpretability that Carta et al.'s rule-based explanations provide.

On the modelling side, only Carta et al. [1] deliberately choose an inherently interpretable model — a shallow Decision Tree — over higher-capacity alternatives, explicitly trading predictive power for explainability. Sami et al. [2] and Bacco et al. [3] both settle on LSTM-family architectures: Sami et al. find Bi-LSTM to be the strongest of eight candidate models (including simpler regressors such as SVR and XGBoost), while Bacco et al. use a plain LSTM throughout but vary its inputs and training regime extensively. Neither of the two deep-learning-centred papers builds in an explainability mechanism comparable to Carta et al.'s rule extraction, nor does either apply a post-hoc explanation tool such as SHAP or LIME — a gap consistent with the wider finding that post-hoc explanation of recurrent architectures in finance remains comparatively rare even as such architectures dominate reported accuracy [5]. It is also notable that none of the three papers adopts a transformer-based architecture, despite the growing use of transformers elsewhere in stock forecasting [6], [9]; whether a self-attention model would outperform the recurrent architectures used here, and whether it could be made more interpretable through attention-weight visualisation, are both open questions for this specific set of tasks.

VIII. DATASETS AND EXPERIMENTAL SETUP

The three studies rely on quite different data sources and validation schemes. Carta et al. [1] combine the large Dow Jones DNA news corpus (2005–2018) with S&P 500 price data across three broad sectors, using a walk-forward scheme with ten successive training/validation/test splits to respect the temporal ordering inherent to financial data. Sami et al. [2] gather price data for ten hand-picked stocks — screened by EPS, current ratio, and beta — from investing.com, Investopedia, and macro trends.net, training on close to a decade of daily data (2010–2020) and deliberately testing on the sharply volatile COVID-19 window of March–June 2020, with a further validation window used to check for over-fitting. Bacco et al. [3] instead assemble a purpose-built dataset covering a single, highly volatile calendar year (2022) for 26 banks split between the US and EU, paired with a matching Twitter dataset drawn from financial commentators, and evaluate several train/test splits corresponding to different combinations of the five experimental factors described in Section 4.3.

These differences in scale, time span, and market coverage make direct numerical comparison across the three studies difficult: Carta et al. evaluate over more than a decade of data and three sectors, Sami et al. deliberately stress-test a short and anomalous window, and Bacco et al. concentrate on a single asset class — banks — during one year of unusual macroeconomic conditions. For a survey, this diversity is nonetheless useful, since it shows that the underlying ideas — lexicon-based features, engineered price attributes, and sentiment fusion — have each been tested under substantially different market conditions.

IX. PERFORMANCE AND RESULTS

Because the three papers target different prediction problems, formulations, and metrics, their reported numbers are not directly comparable in absolute terms, though each offers evidence that its chosen auxiliary signal adds value beyond price data alone. Carta et al. [1] report Balanced Accuracy in the 0.5–0.6 range for their high/low magnitude classification task, arguing that this represents a meaningful gain over both a random classifier and their own earlier unsupervised baseline, with the improvement reaching statistical significance for the Financial sector in particular. Sami et al. [2] report a considerably higher raw accuracy of 98.14% for their Bi-LSTM model on the SAM-augmented feature set, along with an adjusted R^2 above 99% and a portfolio ROI gain of roughly 8.8 percentage points (55.589% versus 46.76%) attributable to the new price feature; this task — price-level regression judged within a fixed error tolerance — is nonetheless fundamentally different from, and not directly comparable to, the classification tasks in [1] and [3]. Bacco et al. [3] report weighted F1-scores averaging 0.68 for US banks and 0.447 for EU banks, underlining that the benefit of sentiment and other auxiliary signals depends heavily on the market rather than holding universally.

A theme common to all three papers is that no single auxiliary signal helps uniformly: the lexicon features in [1] and the sentiment features in [3] both show effectiveness that varies by market or sector, while the PCA and LDA features in [2] turn out to be statistically insignificant and are dropped in favour of the simpler SAM/average feature. This points to the value of any given auxiliary signal being conditional both on the specific market segment and on how the signal is folded into the model — a point reinforced by the more tightly integrated fusion architectures discussed in Section 3, which report consistent gains from sentiment where a simpler concatenation approach, as in [3], does not.

X. ADVANTAGES AND LIMITATIONS

A. Advantages

- 1) Carta et al. [1]: yields genuinely interpretable, rule-based explanations alongside its predictions, speaking directly to the accountability concerns that matter in financial decision-making.

- 2) Sami et al. [2]: validates each engineered feature rigorously through chi-square testing before use, and carries the pipeline all the way from forecasting to a concrete portfolio with measured ROI and Sharpe ratio.
- 3) Bacco et al. [3]: is the only one of the three to compare geographically distinct markets explicitly and to isolate the individual and combined effects of five separate experimental factors.

B. Limitations

- 1) Carta et al. [1]: the lexicon-based features are bag-of-words in character and miss word order, syntax, and deeper semantic relationships between terms; the method has only been tested on the S&P 500 and not on other markets.
- 2) Sami et al. [2]: leaves out news, macroeconomic, and investor-behaviour factors; favours accuracy over computational efficiency, limiting its use for very short-horizon trading; and evaluates the portfolio over only a two-month window.
- 3) Bacco et al. [3]: is confined to a single asset class — bank stocks; its Twitter sentiment corpus is drawn from a curated set of 34 influential accounts and may not reflect broader public sentiment; and it frames the task as binary trend classification rather than continuous price estimation.

XI. CROSS-DOMAIN CASE STUDY: ATTENTION-BASED SIMILARITY BEYOND FINANCE

A theme that runs through the finance literature reviewed above is that a base predictive model can be strengthened by drawing on information from related items — related news articles in [1], related historical price patterns implicitly captured through SAM averaging in [2], or related public commentary in [3]. In every case, though, “relatedness” is defined through a fixed, hand-built rule: percentage lexicon overlap in [1], a simple arithmetic average in [2], or retweet/like-weighted sentiment averaging in [3].

Working in the unrelated field of legislative vote prediction, Wang et al. [4] propose an alternative worth flagging as a direction for future finance research: instead of a fixed similarity formula, they train an attention mechanism to learn directly from data which other legislative bills most closely resemble the one under prediction, then feed a weighted combination of the encoded features of the most similar bills (found via BERT embeddings) into a classifier that also draws on social-relationship features — co-sponsorship, party affiliation, committee membership, and past voting consistency — and on personal legislator attributes. This attention-based similarity mechanism lifts prediction accuracy from 0.7169 (using every other factor but no bill-similarity information) to 0.7838, a 6.69% relative gain, and beats both cosine similarity (0.7566) and Euclidean distance (0.7447) as fixed alternatives for the same task. The authors further show that only a small number of most-similar items — around ten out of hundreds of candidates — is needed to obtain this improvement, and that the bills identified as similar do in fact share consistent voting outcomes, which supports the underlying claim that similarity-weighted auxiliary information carries predictive value.

Although obtained outside finance, this result speaks directly to the literature surveyed here: it suggests that the fixed lexicon-matching approach of Carta et al. [1] or the plain statistical averaging used by Sami et al. [2] might be improved by swapping in a learned, attention-based mechanism capable of adapting to the data instead of assuming a fixed notion of relatedness. This is offered as a suggestive analogy across domains rather than as an established finding for stock forecasting, since none of the core finance papers reviewed here has tested this specific idea; it is revisited as a concrete future direction in Section 14.

XII. RESEARCH CONTRIBUTION AND SCOPE OF THE SURVEY

This survey does not propose or experimentally validate a new forecasting algorithm; its contribution is as a synthesis and comparative analysis of existing published work, together with the identification of research gaps supported by that synthesis. Three points define its scope and its claim to novelty relative to the broader stock-forecasting survey literature.

First, unlike deep-learning-focused surveys that catalogue architectures across dozens of papers at a relatively high level of abstraction [6], [11], this survey performs a close, side-by-side reading of three specific studies chosen precisely because each represents a different strategy for supplementing price data — curated news lexicons, statistically engineered price features, and social-media sentiment — while sharing a common underlying goal. This narrower but deeper focus allows the comparative analysis in Sections 5 and 6 to identify methodological similarities and contrasts (in feature validation, model choice, and explainability) that a broader architectural survey would not surface.

Second, the survey brings together explainable, statistically-driven, and sentiment-driven approaches within a single comparative frame, rather than treating explainability-focused work [1] and accuracy-focused deep learning work [2], [3] as belonging to separate literatures, as they are often discussed in isolation. Reading them together makes visible a trade-off — between explainability and reported predictive performance — that is implicit but not stated in any of the three original papers.

Third, the cross-domain case study in Section 11 is a genuine, if modest, methodological contribution of this survey rather than of the underlying papers: none of the three finance papers considers attention-based similarity learning, and connecting Wang et al.'s [4] legislative-prediction result to the fixed-similarity techniques used throughout the finance literature reviewed here is offered as a concrete, testable direction for future work rather than as a validated finding.

This survey does not claim that the three focal papers are collectively representative of the entire stock-forecasting literature, nor that its expanded literature review in Section 3 is exhaustive; the additional sources cited there were identified through targeted, non-systematic literature search rather than a formal systematic review protocol, and this limitation is stated explicitly rather than implied.

XIII. RESEARCH GAPS

- 1) None of the three finance papers combines all three forms of auxiliary information — curated news lexicons, statistically or dimensionality-reduction-derived price features, and social-media sentiment — within one model, leaving open whether such a combination would complement or duplicate each other.
- 2) Model explainability is addressed directly only by Carta et al. [1]; the higher-accuracy deep models in [2] and [3] (Bi-LSTM, LSTM) stay comparatively opaque, and none of the three papers applies established black-box explanation tools — attention weights, SHAP, or LIME, the latter two mentioned only in passing in [2] — to their deep models. This gap is consistent with the wider literature, which finds that post-hoc explanation of recurrent financial forecasting models remains comparatively rare relative to their reported predictive dominance [5].
- 3) None of the three papers brings in macroeconomic indicators, sector-wide news, or cross-asset information such as commodities or interest rates as auxiliary features, even though both Sami et al. [2] and Bacco et al. [3] explicitly flag this as unaddressed.
- 4) Cross-market generalisability is examined explicitly only by Bacco et al. [3] (US versus EU), and even then only within banking; it is unclear whether the lexicon-based approach of [1] or the SAM-based feature engineering of [2] would transfer to non-US markets or to less liquid asset classes.
- 5) The mechanisms used to combine auxiliary information with base features are fixed and hand-specified across all three finance papers, unlike the learned, attention-based approach shown in the unrelated legislative-prediction setting [4], which there outperformed fixed similarity measures such as cosine similarity and Euclidean distance.
- 6) None of the three papers evaluates a transformer-based architecture, despite transformers' growing use elsewhere in stock forecasting [6], [9]; it is unknown whether a self-attention model would outperform the recurrent architectures used in [2] and [3] on these specific tasks, or whether it would offer any offsetting gain in interpretability through attention-weight visualisation.
- 7) Sentiment is fused with price data through simple concatenation in Bacco et al. [3], whereas more recent hybrid architectures fuse textual and numerical signals through learned attention mechanisms [8], [9]; whether a tighter fusion mechanism would close some of the region-dependent performance gap Bacco et al. report between US and EU banks has not been tested.
- 8) None of the three papers reports a quantitative evaluation of explanation quality itself (for example, fidelity or stability of the extracted rules or feature attributions), reflecting a broader shortage of standardised explainability-evaluation metrics noted in the wider XAI-in-finance literature [5].

XIV. FUTURE RESEARCH DIRECTIONS

- 1) Test whether learned, attention-based similarity mechanisms — as shown outside finance in [4] — could improve on the fixed lexicon-matching approach of [1] or the plain statistical averaging behind the SAM price feature in [2], for example by learning which historical news items, trading days, or comparable companies matter most to a given prediction.
- 2) Bring semantic embeddings, topic models, or frame semantics into lexicon-based methods to move past the bag-of-words limitation that Carta et al. [1] themselves identify, potentially in combination with the sentiment-extraction pipeline already validated by Bacco et al. [3].
- 3) Extend Sami et al.'s [2] feature-engineering approach to macroeconomic, commodity, or cross-asset indicators, and check whether such features remain statistically significant under the same chi-square testing protocol.
- 4) Widen Bacco et al.'s [3] sentiment analysis beyond a curated set of influential Twitter accounts to a more representative sample of public opinion, and extend the US/EU comparison to further markets and asset classes beyond banking.

- 5) Apply post-hoc explainability tools — attention-weight visualisation, SHAP, or LIME — to the higher-accuracy Bi-LSTM and LSTM models in [2] and [3], closing the interpretability gap with the Decision Tree approach in [1] without giving up predictive accuracy, and report a quantitative measure of explanation fidelity where possible.
- 6) Assess computational efficiency and the feasibility of short-horizon, intraday prediction, a dimension Sami et al. [2] explicitly set aside in favour of accuracy.
- 7) Evaluate transformer-based sequence models directly against the LSTM/Bi-LSTM baselines used in [2] and [3] on the same tasks and datasets, to establish whether reported transformer gains elsewhere in the literature [6], [9] carry over to these specific forecasting problems.
- 8) Investigate tighter, learned sentiment-fusion mechanisms — attention-weighted combination rather than simple feature concatenation, as in [8] and [9] — as a way of testing whether the region-dependent benefit of sentiment reported in [3] is a property of the underlying signal or an artefact of the fusion method used.

XV. CONCLUSION

This survey has brought together three recent and complementary approaches to machine-learning-based stock market forecasting: an explainable, lexicon-based Decision Tree method [1]; a feature-engineering and portfolio-optimisation pipeline built around PCA-, LDA-, and statistical-average-derived price attributes combined with Bi-LSTM forecasting [2]; and an LSTM model enriched with FinBERT-derived Twitter sentiment, evaluated across US and EU banking markets under conditions of high macroeconomic uncertainty [3]. Together, these studies show that auxiliary information — textual, statistical, or sentiment-based — can meaningfully improve on price-only forecasting, though its value is neither universal nor guaranteed: engineered features can fail statistical significance tests [2], lexicon-based signals can transfer unevenly across sectors [1], and sentiment can help in one market while offering little in another [3].

Reading the three papers against the broader and more recent literature reviewed in Section 3 sharpens this picture in two ways. First, the field has largely moved, or is moving, toward transformer-based and more tightly fused hybrid architectures that none of the three focal papers evaluates, suggesting that some of their reported limitations — particularly the uneven benefit of sentiment reported by Bacco et al. [3] — may be as much a function of architectural and fusion choices as of the underlying value of the auxiliary signal itself. Second, the systematic XAI-in-finance literature confirms that the trade-off between explainability and predictive accuracy visible across these three papers is not idiosyncratic to them but reflects a broader pattern in the field, where post-hoc explanation of high-accuracy deep models remains comparatively underdeveloped [5].

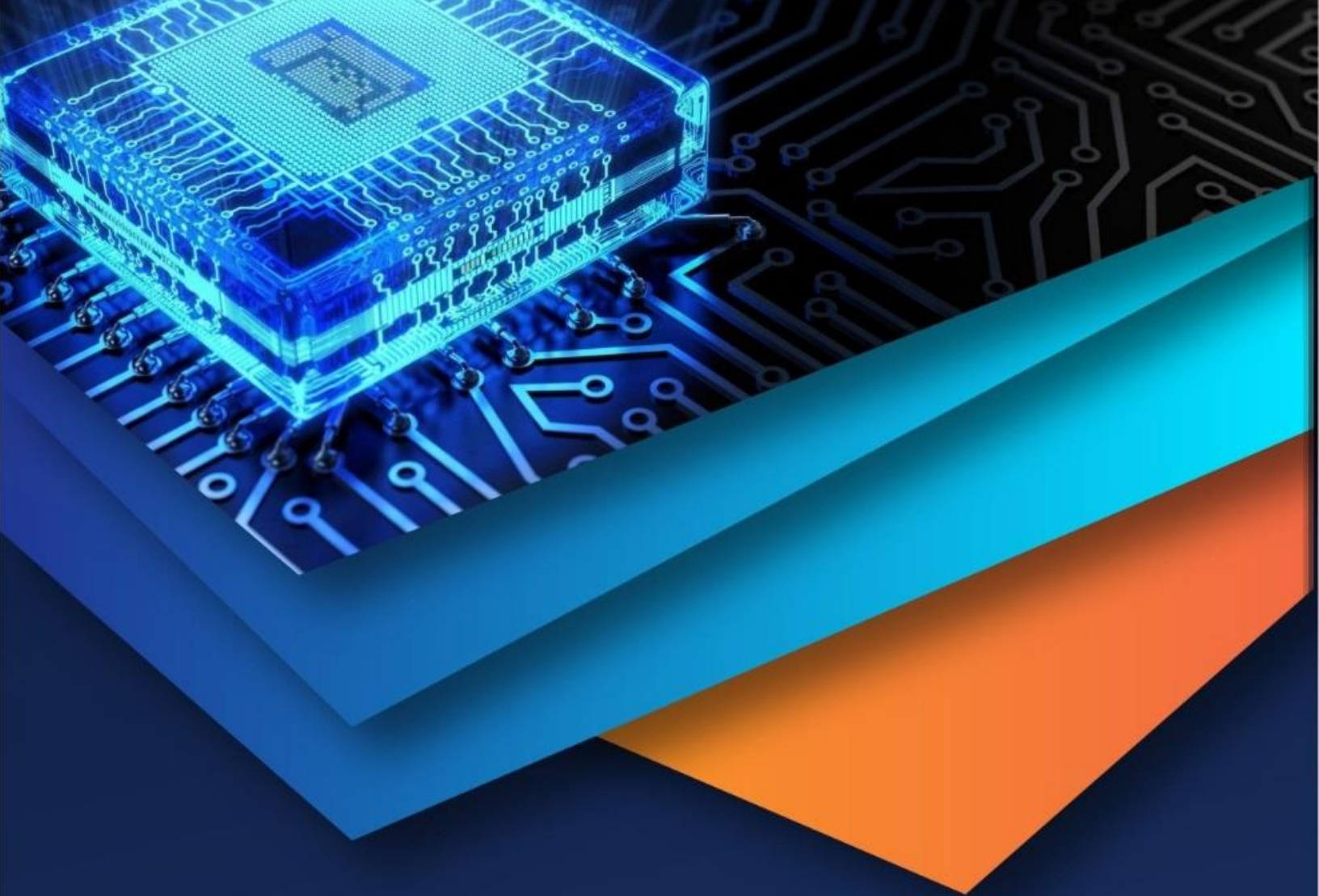
A short cross-domain case study of an attention-based legislative-prediction model [4] further suggests that replacing the fixed, hand-specified similarity and aggregation formulas used throughout this finance literature with a learned, data-driven similarity mechanism is a promising direction that remains largely unexplored here. Future work that combines richer, explainable, and adaptively weighted auxiliary information — rather than treating lexicon-, feature-, and sentiment-based approaches as mutually exclusive choices — appears to be the most promising route toward stock market forecasting systems that are more accurate, more generalisable, and more trustworthy. This survey does not claim to resolve that trade-off, nor to guarantee that any of the directions proposed in Section 14 will succeed; it aims only to state the evidence for these gaps as precisely as the surveyed literature allows.

REFERENCES

- [1] S. M. Carta, S. Consoli, L. Piras, A. S. Podda, and D. Reforgiato Recupero, “Explainable machine learning exploiting news and domain-specific lexicon for stock market forecasting,” *IEEE Access*, vol. 9, pp. 30193–30205, 2021.
- [2] H. M. Sami, M. S. Jalal, M. F. Kabir, and M. N. Huda, “IntPort: An intelligent portfolio construction technique based on financial forecasting by statistical average method,” *IEEE Access*, vol. 13, pp. 35355–35375, 2025.
- [3] L. Bacco, L. Petrosino, D. Arganese, L. Vollero, M. Papi, and M. Merone, “Investigating stock prediction using LSTM networks and sentiment analysis of tweets under high uncertainty: A case study of North American and European banks,” *IEEE Access*, vol. 12, pp. 122239–122248, 2024.
- [4] B. Wang, Y. Li, and C. Xu, “Exploring bill similarity with attention mechanism for enhanced legislative prediction,” *Journal of Social Computing*, vol. 6, no. 2, pp. 112–125, 2025.
- [5] J. Černevičienė and A. Kabašinskas, “Explainable artificial intelligence (XAI) in finance: a systematic literature review,” *Artificial Intelligence Review*, vol. 57, no. 8, article 216, 2024.
- [6] J. Zou, Q. Zhao, Y. Jiao, H. Cao, Y. Liu, Q. Yan, E. Abbasnejad, L. Liu, and J. Q. Shi, “Stock market prediction via deep learning techniques: A survey,” *arXiv preprint arXiv:2212.12717*, 2022 (revised 2023).
- [7] D. Araci, “FinBERT: Financial sentiment analysis with pre-trained language models,” *arXiv preprint arXiv:1908.10063*, 2019.
- [8] W. Lu, J. Li, J. Wang, R. Zhang, and Y. Qin, “A CNN-BiLSTM-AM method for stock price prediction,” *Neural Computing and Applications*, vol. 33, pp. 4741–4753, 2021.



- [9] Y. Nejatbakhsh and M. Aliasgari, "Enhancing stock market prediction with hybrid deep learning: Integrating LSTM, transformer attention, federated learning, and sentiment analysis," *IEEE Access*, vol. 14, pp. 3926–3942, 2025.
- [10] P. Sonkiya, V. Bajpai, and A. Bansal, "Stock price prediction using BERT and GAN," *arXiv preprint arXiv:2107.09055*, 2021.
- [11] W. Jiang, "Applications of deep learning in stock market prediction: Recent progress," *Expert Systems with Applications*, vol. 184, article 115537, 2021.
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019, pp. 4171–4186.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)