

Artificial Intelligence in Stock Market Prediction: A Comparative Case Study of Machine Learning and Deep Learning Approaches

Anurag Ajay Satam*, Dr. Manoj Singh**

*(Roll No: SMCS2627114, Department of Computer Science, SIES College of Arts, Science and Commerce, Mumbai, India

Email: Anuragfms2526114@asc.sies.edu.in)

** (Department of Computer Science, SIES College of Arts, Science and Commerce, Mumbai, India

Email: manojks@sies.edu.in)

Abstract:

Predicting stock prices and market direction is a long-standing problem in finance and computer science, and artificial intelligence (AI) is widely proposed as a way to solve it. This paper presents a comparative case study of six recent publications on AI-based stock market prediction, covering classical machine learning, deep learning, metaheuristic hyperparameter optimization, and sentiment analysis. After reviewing the market-theory background and the standard prediction pipeline, we re-examine the reported results of four empirical studies: linear regression, LSTM, and nine classifiers on Apple stock and tweet sentiment; seven classifiers on seven G7 stock indices; ANN and CNN models on India's NIFTY index; and genetic-algorithm and artificial-rabbits-optimized LSTMs on the 30 Dow Jones stocks. Summary statistics recomputed from the published tables show that (i) the best algorithm depends on the market, with ANN best on four of seven G7 indices and logistic regression on the other three; (ii) metaheuristic-tuned LSTMs cut mean MAPE from 3.9–4.5% to about 2.0% relative to untuned networks, yet the two optimizers differ by only 0.04 percentage points; and (iii) headline accuracies ranging from roughly 50% to nearly 99% are not comparable, because the studies use different targets and metrics. We conclude that AI methods capture useful short-horizon structure, but none of the reviewed evidence demonstrates reliable, cost-adjusted outperformance of the market.

Keywords — Stock Market Prediction, Machine Learning, Deep Learning, LSTM, Convolutional Neural Network, Sentiment Analysis, Efficient Market Hypothesis, Technical Analysis.

I. INTRODUCTION

Equity markets channel capital between investors and firms, and the ability to anticipate price movements has obvious value to investors, firms, and regulators [3], [5]. Prices, however, respond to macroeconomic conditions, company results, news, and investor sentiment, which makes forecasting difficult and, for some economists, impossible [1], [5]. Since the 1990s, researchers have applied artificial intelligence to the problem, and the resulting literature has grown quickly: a systematic review found 2,326 Scopus-indexed papers on AI for stock-market investment published between

1995 and 2019, with the annual count rising exponentially [3].

Automation is also increasingly common in practice. The review in [3] cites a report that computers already execute the vast majority of hedge-fund trades, although about 90% of those trades still follow hard-coded procedures rather than learned models. At the same time, the strongest claims in the literature—accuracies above 97% or 98%—sit alongside sober conclusions that current AI is not yet reliable enough to beat the market [1], [4]. These statements are hard to reconcile without looking closely at what was measured.

This paper therefore synthesizes six recent studies [1]–[6] into a single comparative case study. Sections II and III provide related work and background on market theory and the prediction pipeline. Section IV summarizes the techniques used, and Section V discusses evaluation metrics and the difficulty of comparing results. Sections VI–IX examine four empirical case studies in detail, Section X reviews sentiment-analysis and hybrid approaches, Section XI discusses what the evidence does and does not support, and Section XII concludes.

II. RELATED WORK

Mokhtari, Yen, and Liu [1] evaluate machine learning for both technical analysis (regression on price indicators for Apple stock) and fundamental analysis (classification of tweet sentiment), and conclude that AI is not yet accurate enough to beat the market. Rajesh, Manjusha, and Thippanna [2] outline a four-stage pipeline of data collection, cleaning, model training, and next-day prediction using linear regression, decision tree, and random forest models, and compare the strengths and limitations of six algorithms qualitatively; the paper reports no numerical results. Ferreira, Gandomi, and Cardoso [3] survey 2,326 papers and divide them into portfolio optimization, price prediction, sentiment analysis, and combinations of these. Mukherjee et al. [4] predict the NIFTY index with a backpropagation ANN and with a CNN that consumes two-dimensional histograms, and test the CNN through the COVID-19 crash. Ayyildiz and Iskenderoglu [5] compare seven classifiers on the daily direction of seven G7 stock indices. Gülmez [6] optimizes the architecture of LSTM networks with the genetic algorithm and the recently proposed artificial rabbits optimization (ARO) algorithm and evaluates them on the 30 Dow Jones Industrial Average (DJIA) stocks. Together, these studies span survey, method, and empirical evidence across three continents' markets.

III. FOUNDATIONS: MARKET THEORY AND THE PREDICTION PIPELINE

A. Efficient and Adaptive Market Hypotheses

The Efficient Market Hypothesis (EMH) holds that prices fully reflect available information, so

price changes are unpredictable and past prices cannot be used to earn excess returns [1], [3], [5]. The Adaptive Market Hypothesis (AMH) relaxes this view: markets evolve and alternate between periods of efficiency and inefficiency, so exploitable patterns may exist at certain times [1], [5]. All six studies implicitly adopt the AMH position, since each treats historical or textual data as informative about future prices.

B. Technical and Fundamental Analysis

Technical analysis predicts prices from historical prices, volumes, and indicators derived from them, whereas fundamental analysis uses company financial statements, economic indicators, and, in modern practice, news and social-media sentiment [1]–[3]. Commonly used technical indicators include the simple and exponential moving averages (SMA, EMA), the relative strength index (RSI), moving average convergence/divergence (MACD), on-balance volume (OBV), momentum, stochastic oscillators, and the commodity channel index [1], [5]. The EMA and RSI are defined in [1] as

$$EMA_t = \Gamma \cdot (CP_t - EMA_{t-1}) + EMA_{t-1}, \quad \Gamma = \frac{2}{\Delta + 1} \quad (1)$$

$$RSI = 100 - \frac{100}{1 + RS}, \quad RS = \frac{AvgGain}{AvgLoss} \quad (2)$$

where CP is the closing price, Δ the averaging period, and RS the ratio of average gains to average losses. Features are usually rescaled before training; [1] applies min–max scaling,

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3)$$

C. The Prediction Pipeline

Despite differences in detail, the studies follow a common pipeline (Fig. 1): acquire data (prices from Yahoo Finance or exchange websites; tweets or news for sentiment), clean and reduce it (missing values, duplicates, outliers, correlation-based selection, PCA), engineer features, train one or more models, tune hyperparameters on validation data, and evaluate on held-out data [1]–[3]. The output is either a numeric forecast (regression) or a rise/fall/hold signal (classification). The review in [1] reports that over 90% of stock-prediction algorithms in the literature are classifiers, and [3]

notes that regression is typically used for prices and classification for trends.

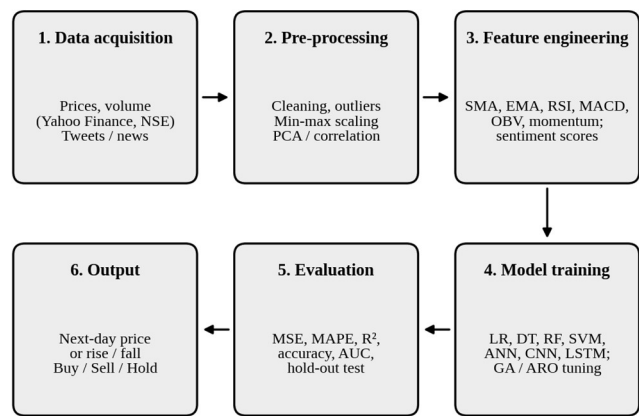


Fig. 1 Common AI stock-prediction pipeline synthesized from [1]–[3].

IV. MACHINE LEARNING AND DEEP LEARNING TECHNIQUES

A. Classical Machine Learning

Linear regression fits a hyperplane to the features and is easy to interpret; logistic regression applies the same idea to binary rise/fall targets [1], [2]. Decision trees split the feature space with learned rules and can overfit unless constrained, while random forests average many decorrelated trees to improve stability at the cost of speed [2]. Support vector machines find a maximum-margin separator between up and down days, k-nearest neighbors classify by similarity to stored examples, and naive Bayes applies Bayes’ theorem under an independence assumption [1], [5]. Table I summarizes the strengths and limitations reported across the studies.

TABLE I
STRENGTHS AND LIMITATIONS OF PREDICTION TECHNIQUES

Technique	Strengths	Limitations	Ref.
Linear / logistic regression	Simple, interpretable; suits steady trends	Poor at sudden changes	[2]
Decision tree	Handles non-linear data; easy to visualize	Overfits if not controlled	[2]
Random forest	More accurate and stable than one tree	Slower than single models	[2]
SVM	Strong for up/down classification	Not ideal for large datasets	[2]
ANN (backprop.)	Models complex non-linear relations	10,000+ epochs; some overfitting	[4]
CNN (2-D histogram)	Fewer epochs and less training time than ANN	Extra cost of building images	[4]

Technique	Strengths	Limitations	Ref.
LSTM	Excellent for time series	Needs more data and training time	[2]
ARIMA	Classical time-series forecasting	Less flexible than ML	[2]
GA / ARO-tuned LSTM	Automated architecture and hyperparameter search	About 8–9 days of search for 30 stocks	[6]

Sources: [2], [4], [6]. ANN = artificial neural network; CNN = convolutional neural network.

B. Neural and Deep Learning Models

An ANN combines weighted inputs and a bias through an activation function and learns by backpropagating the error gradient, $\Delta w = -\eta \partial E / \partial w$ [4], [6]. Convolutional neural networks (CNNs), designed for images, were adapted in [4] by converting a price series into two-dimensional histograms. Long short-term memory (LSTM) networks use input, forget, and output gates to retain information over long sequences and are widely used for time series [2], [6]. The review in [3] identifies a marked recent increase in deep-learning methods, but also notes that little work has isolated the contribution of individual input features.

C. Metaheuristic Hyperparameter Optimization

LSTM performance depends on hyperparameters such as the number of layers and neurons, dropout rate, optimizer, and learning rate. In [6] these are treated as variables in a search space (1–20 neurons per layer, optional second and third LSTM layers, dropout 0.3–0.7, five optimizers, and learning rates from 10^{-2} to 10^{-6}) and searched with a genetic algorithm (GA) and with ARO, both run with 50 individuals for 50 iterations. ARO alternates between “detour foraging” (exploration) and “random hiding” (exploitation), switching according to an energy factor that decays over the iterations [6].

V. EVALUATION METRICS AND STUDY OVERVIEW

Regression studies report the mean absolute percentage error (MAPE), root mean squared error (RMSE), and coefficient of determination (R^2) [1], [6]:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{y_i} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

Classification studies report accuracy, precision, recall, F1, and the area under the ROC curve (AUC) [1], [5]. The NIFTY study [4] instead defines the “accuracy” of a price forecast t with output o as

$$Accuracy = (1 - \frac{|t - o|}{t}) \times 100 \quad (7)$$

Averaged over a test set, this quantity equals 100 minus the MAPE (in percent), which allows results reported in different forms to be placed on a common scale in Section XI. Table II lists the six studies with their data, methods, and headline results as reported.

TABLE II
OVERVIEW OF THE SIX STUDIES

Ref.	Data and task	Methods	Headline result (as reported)
[1]	AAPL 2010–2021 (price); ~6,000 tweets (sentiment)	LR, LSTM; 9 classifiers	LR RMSE 1.82; SVM accuracy 75.5%
[2]	5 years of prices; next-day price	LR, DT, RF	Qualitative; no metrics reported
[3]	2,326 Scopus papers, 1995–2019	Systematic review	Combined approaches reported superior; few studies
[4]	NIFTY, Apr 2008–Apr 2018; COVID period 2020	ANN, CNN (2-D histograms)	97.66% (ANN), 98.92% (CNN), 91% in COVID case
[5]	7 G7 indices, daily 2012–2021; next-day direction	7 classifiers, 10 indicators	ANN best on average; best single result 93.5% (FTSE 100)
[6]	30 DJIA stocks, 2018–2023; next-day price	ANN, 3 LSTMs, LSTM-GA, LSTM-ARO	LSTM-ARO lowest MSE on 18 of 30 stocks

LR = linear regression; DT = decision tree; RF = random forest.

VI. CASE STUDY I: TECHNICAL AND SENTIMENT ANALYSIS ON APPLE

Mokhtari et al. [1] built two separate predictors for Apple (AAPL). For technical analysis they downloaded more than ten years of daily data (2010–2021) from Yahoo Finance, formed 60 features including prices, moving averages, MACD, and RSI, selected the features most correlated with the target closing price, merged redundant ones, applied min–max scaling, and split the data into

training, validation, and test sets. Linear regression and an LSTM were compared (Table III).

TABLE III
PRICE REGRESSION ON AAPL: LINEAR REGRESSION VS. LSTM [1]

Metric	Linear regression	LSTM
R^2	1.00	0.99
Explained variation	1.00	0.99
MAPE	1.56	2.99
RMSE	1.82	3.42
MAE	1.18	2.30

Source: [1].

Linear regression outperformed the LSTM on every metric. An R^2 of 1.00, however, should be read with care: the feature set contains same-day open, high, and low prices, which are not available before the close, so the task resembles interpolation of the closing price more than forecasting. The authors themselves note that the model cannot precisely predict the value for the next business day and is therefore not sufficient for long-term investment [1].

For fundamental analysis, [1] labeled nearly 6,000 tweets about Apple as positive or negative, reduced the features with PCA, and trained nine classifiers (Table IV). The SVM was best with 75.5% accuracy and an AUC of 0.76; logistic regression and random forest reached 72.7%; the decision tree was worst at 62.0%. The ROC legend in [1] lists slightly different AUC values for some models (e.g., 0.74 for the SVM), but the ranking is unchanged. The authors judge this accuracy insufficient for trading and warn that it could lead to capital loss.

TABLE IV
TWEET-SENTIMENT CLASSIFICATION FOR APPLE [1]

Classifier	Accuracy	F1-score	AUC
Logistic regression	0.727	0.726	0.73
Gaussian NB	0.634	0.632	0.63
Bernoulli NB	0.644	0.644	0.64
Decision tree	0.620	0.620	0.62
Random forest	0.727	0.727	0.73
KNN	0.684	0.684	0.68
SVM	0.755	0.755	0.76
XGBoost	0.709	0.709	0.71
ANN	0.684	0.684	0.68

Source: [1]. Best value in bold. NB = naive Bayes; KNN = k-nearest neighbors.

VII. CASE STUDY II: DIRECTION PREDICTION ON G7 STOCK INDICES

Ayyildiz and Iskenderoglu [5] predicted whether the next day's close would rise or fall for the main index of each G7 country: NYSE 100, NIKKEI 225, FTSE 100, CAC 40, DAX 30, FTSE MIB, and TSX. Daily data from 1 January 2012 to 31 December 2021 were obtained from Investing.com, and ten technical indicators (Section III-B) were computed as inputs. Seven classifiers were trained on an 80:20 train/test split, chosen after comparing splits from 50:50 to 90:10 [5].

TABLE V
NEXT-DAY DIRECTION ACCURACY ON G7 INDICES [5]

Index	DT	RF	KNN	NB	LR	SVM	ANN
NYSE100	0.534	0.641	0.530	0.641	0.821	0.722	0.837
NIKKEI225	0.541	0.572	0.485	0.590	0.816	0.800	0.810
FTSE100	0.589	0.593	0.512	0.607	0.850	0.816	0.935
CAC40	0.596	0.627	0.510	0.641	0.836	0.810	0.830
DAX30	0.547	0.506	0.496	0.623	0.808	0.816	0.818
FTSE MIB	0.566	0.628	0.483	0.638	0.822	0.800	0.851
TSX	0.592	0.564	0.544	0.643	0.865	0.847	0.859
Mean	0.566	0.590	0.509	0.626	0.831	0.802	0.849

Source: [5]. Best value per row in bold. The means are recomputed from the per-index values; those printed in [5] differ slightly for KNN, LR, SVM, and ANN (0.503, 0.826, 0.794, 0.834), but the ranking is identical.

Three algorithms—ANN, logistic regression, and SVM—exceeded 70% on every index, whereas decision tree, random forest, naive Bayes, and KNN scored between 48% and 64% (Fig. 2). KNN fell below 50% on three of the seven indices. ANN was best on the NYSE 100, FTSE 100, DAX 30, and FTSE MIB, and logistic regression on the NIKKEI 225, CAC 40, and TSX. The authors attribute the difference to each market's volatility, structure, and economic conditions [5].

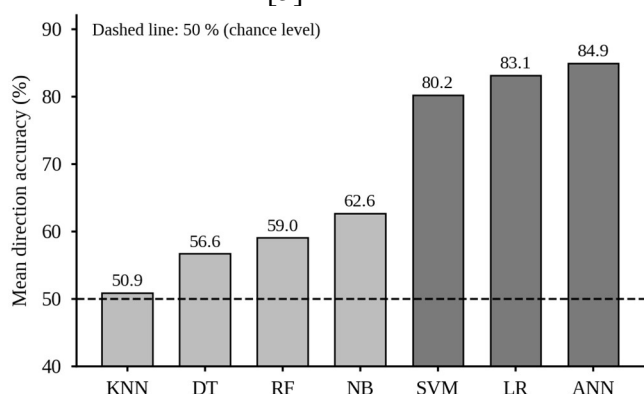


Fig. 2 Mean next-day direction accuracy of seven classifiers across the seven G7 indices (recomputed from [5]).

Two cautions apply. First, apart from the FTSE 100, where ANN leads logistic regression by 8.5 points, the gap between the best and second-best algorithm is at most 2.9 points and as small as 0.2. With roughly 500 observations per index and a 20% test share, each accuracy is estimated from on the order of 100 days, for which the binomial standard error at 80% accuracy is about 4 points; such differences are effectively ties. Second, the reported counts of 495–512 observations per index are much smaller than a ten-year daily series (about 2,500 days) would suggest, and the paper does not explain the difference.

VIII. CASE STUDY III: DEEP LEARNING ON THE NIFTY INDEX

Mukherjee et al. [4] used NIFTY open, high, low, and close prices from the National Stock Exchange of India for April 2008–April 2018. Their backpropagation ANN takes the four prices of the two previous days (eight inputs), passes them through two hidden layers of two sigmoid neurons each, and outputs the four prices for the next day; it was trained on 2,000 input combinations for more than 10,000 epochs at a learning rate of 0.1 and tested on more than 2,500 combinations.

Their second model is a CNN that operates on images. For each price series, 50 pairs of consecutive-day prices (price on day t against day $t+1$) are binned into a 20×20 grey-scale two-dimensional histogram; the network applies convolution, max-pooling, a second convolution and pooling, and a fully connected layer, and outputs a ratio that is multiplied by the previous day's value. It was trained with Adam (learning rate 0.01, MSE loss) on 1,000 examples for about 150 epochs and tested on more than 1,600 matrices. On the authors' dual-core PC with 2 GB of memory, training took about 3 h for the ANN and 1.5 h for the CNN [4].

TABLE VI
NIFTY PREDICTION ACCURACY, ANN VS. CNN [4]

Feature	ANN (%)	CNN (%)
Open	97.87	98.63
High	97.77	98.94
Low	97.53	99.06
Close	97.47	99.03
Overall	97.66	98.92

Source: [4]. Accuracy defined by Eq. (7). Overall values are the means of the four features.

The CNN was more accurate for every feature and needed less data and time. As a stress test, the CNN was applied from 15 January to 5 August 2020, spanning the COVID-19 crash. Average accuracy fell to about 91–94% in the first days of the decline, then recovered, and the reported aggregate over the period was 91% [4]. This is a useful demonstration that accuracy degrades during a regime shift, though the paper's statements of 95–98% recovery and a 91% aggregate are not easily reconciled from the text. The authors compare their results with two earlier works (76.9% for a boosted decision tree and 57.2% for an LSTM variant), but those studies used different tasks and metrics, so the comparison is indicative only.

Because Eq. (7) measures relative error against the actual price, and day-to-day changes in an index are small relative to its level, even a naive forecast that repeats the previous day's value would score highly. The paper does not report such a baseline, so the 97–99% figures show that the models track the level of the index but not, by themselves, that they add information about its next move.

IX. CASE STUDY IV: METAHEURISTIC-OPTIMIZED LSTM ON THE DJIA

Gülmez [6] predicted the next-day closing price of each of the 30 DJIA constituents from the previous 20 days of prices, using five years of Yahoo Finance data (2018–2023) split 80:20. Six models were compared: a two-hidden-layer ANN (20-10-10 with 50% dropout); LSTM1D, LSTM2D, and LSTM3D with one, two, and three LSTM layers of 10 units; and LSTM-GA and LSTM-ARO, whose layer sizes, dropout, optimizer, and learning rate were chosen by the genetic algorithm and by ARO respectively. Results were reported per stock for MSE, MAE, MAPE, and R^2 . Since raw MSE and MAE scale with price level, we summarize the published tables with scale-free measures (Table VII).

TABLE VII
SUMMARY OF 30-STOCK DJIA RESULTS, RECOMPUTED FROM [6]

Model	Mean MAPE (%)	Mean R^2	Stocks with $R^2 < 0$	Best MSE (of 30)	Time per stock (s)
LSTM-ARO	2.00	0.912	0	18	10.1

Model	Mean MAPE (%)	Mean R^2	Stocks with $R^2 < 0$	Best MSE (of 30)	Time per stock (s)
LSTM-GA	2.04	0.911	0	10	9.6
LSTM1D	4.06	0.645	2	2	8.3
LSTM2D	4.30	0.581	2	0	12.5
LSTM3D	4.52	0.559	3	0	16.7
ANN	3.86	0.715	0	0	4.5

Source: computed from Tables 7, 9, 10, and 11 of [6]. The 10.1 s per stock is inference time; the GA and ARO searches themselves took about 8.4 and 8.7 days for all 30 stocks.

Both optimized LSTMs roughly halve the MAPE of the untuned networks and reach a mean R^2 above 0.91, while the untuned LSTMs produced negative R^2 (worse than predicting the mean) for KO and UNH, and LSTM3D also for CVX. LSTM-ARO had the lowest MSE on 18 stocks, LSTM-GA on 10, and LSTM1D on 2 (HD and TRV); both optimized models beat all four untuned models on 28 of the 30 stocks.

The difference between the two optimizers is small: mean MAPE was 2.00% versus 2.04%, mean R^2 0.912 versus 0.911, and the median MSE ratio of ARO to GA was 0.97. The evidence therefore supports the value of automated hyperparameter search rather than the superiority of ARO specifically. Two further points matter. The flowchart in [6] indicates that the optimizer's fitness is computed from prediction error on the test data, and no separate validation split is described, so the tuned models' test errors may be optimistic relative to the fixed-architecture baselines. And the search cost, about 8.4–8.7 days for 30 stocks, is a one-off expense, after which inference for the whole index takes about five minutes [6].

X. SENTIMENT ANALYSIS AND HYBRID MODELS

The survey in [3] found 289 papers on financial sentiment analysis, alongside 1,719 on machine-learning price forecasting and 693 on portfolio optimization. Early sentiment studies used news, which tends to be neutral and hard to classify as good or bad; more recent work favors social-media messages, which are far more numerous but require heavier pre-processing [3]. The results reviewed in [3] are mixed but generally positive. A Twitter mood study found that only one of seven mood dimensions ("calm") correlated with the DJIA and improved prediction. A study of 1.6 million labeled tweets reached close to 80% sentiment accuracy

and found predictive information about price movements, and a study on Apple tweets reached 76.7% movement accuracy. Another, on the SSE 50 index, reported that adding sentiment features raised accuracy from 63–72% to 80–90%.

The survey concludes that combining approaches (for example, SVM classifiers with portfolio optimization, or price features with textual sentiment) outperforms single methods, although few such studies exist [3]. It also reports a case in which a deep neural network was more accurate than an ANN but produced lower simulated trading returns and higher variance, illustrating that classification accuracy and trading performance are different objectives [3]. The Apple tweet results in Case Study I (75.5% for the best classifier) fall within the range reported by these studies [1].

XI. DISCUSSION

A. Rankings Depend on Data and Task

No algorithm was universally best. In [5], ANN had the highest mean accuracy but logistic regression won on three of the seven indices, and [5] cites earlier studies in which random forest or SVM was best on the same or similar indices. In [1], linear regression beat an LSTM at price regression, whereas in [6] tuned LSTMs greatly outperformed an ANN. The differences in targets, features, periods, and validation procedures are large enough that rankings should be treated as conditional on the setting, echoing the review's observation that the contribution of individual features remains understudied [3].

B. Headline Numbers Are Not Comparable

The reported figures include price-level "accuracy" of 97–99% [4], direction accuracy of 50–93% [5], sentiment accuracy of 62–76% [1], and MAPE of about 2% [6]. Since Eq. (7) averages to $100 - \text{MAPE}$, price-forecast results can be placed on one scale (Table VIII). On that scale, the CNN of [4], the linear regression of [1], and the tuned LSTMs of [6] all achieve MAPE of about 1–2%, whereas the untuned networks of [6] are at 3.9–4.5%. Even so, the assets, periods, and inputs differ, and none of the studies reports a persistence baseline against which these errors could be judged. Direction accuracy is a different quantity altogether:

80–85% next-day accuracy on major indices [5] is far above the 50% expected under the EMH, and a single 80:20 split of about 500 observations cannot exclude favorable-sample effects.

TABLE VIII
PRICE-FORECAST ERROR EXPRESSED AS MAPE (%)

Ref.	Model (asset)	MAPE (%)	Source of value
[4]	CNN (NIFTY)	1.08	100 – 98.92
[1]	Linear regression (AAPL)	1.56	Reported
[6]	LSTM-ARO (DJIA mean)	2.00	Computed
[6]	LSTM-GA (DJIA mean)	2.04	Computed
[4]	ANN (NIFTY)	2.34	100 – 97.66
[1]	LSTM (AAPL)	2.99	Reported
[6]	ANN (DJIA mean)	3.86	Computed
[6]	LSTM1D (DJIA mean)	4.06	Computed

Assets, periods, and input windows differ across studies; values indicate order of magnitude only.

C. Missing Evidence on Profitability

Only the survey [3] discusses trading simulations; the four empirical studies [1], [4]–[6] report prediction error or classification accuracy but not returns, transaction costs, or risk-adjusted performance. This matters because an accurate classifier does not necessarily yield profit, as the DNN-versus-ANN example in [3] shows, and [1] warns that classifiers of the accuracy achieved could result in capital loss.

D. Limitations and Future Work

This paper is a synthesis of published results, not a new experiment, and its statistics are only as reliable as the source tables (several of which contain small internal inconsistencies noted above). Studies [5] and [2] themselves acknowledge that macroeconomic variables, additional indicators, longer periods, and ensemble or deep models were not explored, or, in the case of [2], that no quantitative results were reported. Future work should adopt walk-forward validation with a validation set separate from the test set, include naive and buy-and-hold baselines, report cost-adjusted trading performance, and test hybrid models that combine technical indicators, sentiment, and macroeconomic data [1], [3], [5].

XII. CONCLUSION

This paper compared six recent studies of AI-based stock market prediction and re-derived their key statistics. The evidence shows that neural networks, logistic regression, and SVMs can classify next-day direction of major indices at 70–93% on held-out data [5], that CNNs and tuned LSTMs forecast price levels to within about 1–2% [4], [6], and that automated hyperparameter search substantially improves LSTMs, although the choice of GA versus ARO matters little [6]. It also shows that the best model changes with the market, that sentiment classification from tweets remains in the 60–77% range [1], and that performance can degrade during crises [4]. Because the studies use incompatible metrics, omit naive baselines, and do not report trading costs or returns, the claim that AI reliably beats the market is not supported by this body of evidence, consistent with the conclusion of [1]. Robust validation, transparent baselines, and cost-aware backtesting are the most important steps toward a definitive answer.

REFERENCES

- [1] S. Mokhtari, K. K. Yen, and J. Liu, “Effectiveness of Artificial Intelligence in Stock Market Prediction based on Machine Learning,” *International Journal of Computer Applications*, vol. 183, no. 7, pp. 1–8, Jun. 2021.
- [2] G. Rajesh, K. Manjusha, and G. Thippanna, “A Comprehensive Approach for AI-Based Stock Market Prediction,” *Research Digest on Engineering Management and Social Innovations*, vol. 2, no. 4, pp. 510–516, Apr. 2026.
- [3] F. G. D. C. Ferreira, A. H. Gandomi, and R. T. N. Cardoso, “Artificial Intelligence Applied to Stock Market Trading: A Review,” *IEEE Access*, vol. 9, pp. 30898–30917, 2021.
- [4] S. Mukherjee, B. Sadhukhan, N. Sarkar, D. Roy, and S. De, “Stock market prediction using deep learning algorithms,” *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 82–94, 2023.
- [5] N. Ayyildiz and O. Iskenderoglu, “How effective is machine learning in stock market predictions?,” *Heliyon*, vol. 10, art. no. e24123, 2024.
- [6] B. Gülmez, “Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm,” *Expert Systems With Applications*, vol. 227, art. no. 120346, 2023.