IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CE-41 Machine Learning and Deep Learning-Based Classification Methods for Stock Market Prediction A Review

Mahesh M. Mahajan[1], Dr. Nilesh A. Suryawanshi[2]

[1]Research Scholar : 0000-0003-4611-3070
SES's Gangamai COE, Nagaon, Dhule, India, 424001
[2]Assistant Professor : 0009-0001-1771-7003
SES's Gangamai COE, Nagaon, Dhule, India, 424001
maheshmahajan777@gmail.com

Abstract – *Predicting the stock market has always been tricky, given how random, complicated, and unstable prices can be. Classification-based methods—those that forecast whether prices will go up or down—have become popular, mostly because they fit real trading strategies well. This paper takes a close look at classification techniques used in stock market prediction, from old-school machine learning models to ensemble methods and deep learning. We dig into how these approaches work, what they do well, where they fall short, and how suitable they are for analyzing financial time-series data. On top of that, we highlight major research gaps when it comes to temporal stability, feature interaction, and robustness, and point out promising directions for future studies.*

Keywords- Stock Market Prediction, Classification Models, Machine Learning, Deep Learning, Financial Time-Series.

INTRODUCTION

Price movements in the stock market are influenced by various dynamic and complex factor interdependencies including macroeconomic indicators, company-specific fundamentals, market sentiment and global events [1][9]. The non-stationarity (time-varying), high volatility and low signal-to-noise ratio of financial time-series data makes it very difficult to make accurate predictions. In particular, traditional statistical models have had great difficulty capturing the aforementioned complexities, which has driven an increased use of machine learning; and in particular, classification-based machine learning techniques (methods) to predict stock markets [2] [10]. Predicting the exact price of a stock in the future is often less important than determining the direction of the price movement of a stock [3]. Therefore, the prediction problem is essentially a two or multi-class classification problem which seeks to determine whether a stock or index will generate positive, negative or neutral returns over a defined period. This formulation of stock market prediction relates approximately to real-world decision-making processes when implementing algorithmic trading, portfolio allocation or risk management; accordingly, the ability to accurately predict directional price movements will greatly influence profitability [13].

In this regard, classification methods have several advantages. They can deal with high-dimensional feature spaces which results from the significant degree of engineering associated with creating technical indicators such as trend indicators, momentum indicators, volatility indicators and volume-based indicators [14] and modern classifiers including traditional classifiers such as logistic regression and support vector machines through to advanced hierarchical models and deep learning architectures can also represent complex non-linear relationships and interactions between features. Due to their inherent properties, these types of classifiers are ideally suited to describe interactions in complex financial markets characterized by nonlinear dependencies that change with time. At the same time, various difficulties arise in the use of the classifier framework for stock market predictions. First, due to the dynamic nature of financial data, the patterns learned from past observations are often unable to be generalized and applied successfully to future periods due to changes in regime and market characteristics [15]. Another common problem associated with the use of indicators is overfitting when technical indicators have redundant information and are too similar to each other. Finally, an important problem is related to the incorrect processing of time series data [4], which leads to the introduction of data leakage and optimistic estimates of results. In order to solve these problems, current research focuses on creating stable and adaptable models based on feature selection, temporal validation techniques, and modeling of interactions [6].

Of particular interest in this regard are approaches that prioritize stability and consistency in feature selection processes, as well as taking into account the dependencies between indicators used.

This paper offers an exhaustive analysis of classification techniques employed for stock market forecasting, evaluating their theoretical underpinnings, practical applications, and empirical efficacy [8]. It also points out major flaws in current methods and suggests new areas of research, focusing on stability-aware and interaction-driven methods that are necessary for improving the state of the art in financial machine learning [11] [12].

LITERATURE REVIEW

The use of classification methods for stock market prediction has come a long way in the last two decades from using conventional statistical and machine-learning methods to using more sophisticated deep and combined approaches. We provide a literature survey along the same lines.

Traditional Machine Learning Based Classification

Initial studies of stock price trend prediction utilized common classifier methods such as Logistic Regression, Decision Tree and Naïve Bayes. These methods had been chosen primarily since they were easier to use. A groundbreaking study by Patel et al., found that Decision Tree, Naive Bayes and Artificial Neural Network classifiers could predict the stock trend based on technical indicators data [3] but, as the models made linear assumptions and were based on historical data without using any proper time series modeling, their performance was limited. Studies show that for developing a baseline classifier, Logistic Regression is used but it didn’t capture the dependencies of data. Kumbure et al. performed empirical studies on more than 100+ research articles and after studying they show that Traditional models are still chosen since it is computationally faster and easy to interpret, especially for low-dimensional data [29]. However they suffer from: Poor handling of non-linear relationships Limitations in modeling temporal dependencies they are also sensitive to feature scaling and noise.

Support vector machines (SVM)

Support vector machines (SVM) are used widely classifying data when the decision boundary between classes is not linear and relies on kernel functions to deal with that. Machine learning techniques are becoming more acceptable in financial prediction tasks, SVM among them [2] [22]. They discovered that when using various technical indicators these methods outperformed several traditional ML models (eg. random forest).
Among their disadvantages they fall into the following categories:
Computationally intensive when applied to large datasets.
Varies significantly depending on the kernel and hyperparameter choices not easily scalable for online use.

Ensemble methods

Combining models to improve predictive power has gained prominence with Random Forest and Gradient Boosting approaches leading the way. In a recent review, they have shown ensemble methods outperforming single classifiers consistently and improving generalisation rates [5] [12]. These methods combine the outputs of several base estimators in order to improve general reliability and robustness. Random forests work effectively with many variables financial data and typically have low variance. Other work shows that there are many ensemble models being applied to these markets gradient boosting, adaboost xgboost [18] etc. As these models have no awareness of the temporal evolution of the data, obtaining many accurate predictions can be more difficult as there may be no obvious causal relation evident in the chosen feature set.

Artificial Neural Networks (ANN's)

Artificial Neural Networks (ANN's) were one of the first non-linear models for prediction in Stock Markets. Ji et al assessed the effectiveness of ANN and SVM in predicting Stock Market movements and announced that ANN's were better at capturing non-linear relationships allowing for greater accuracy when classifying certain datasets [25]. In addition, earlier works illustrated that ANN's can identify the hidden patterns in Time-Series Financial Data through the use of many hidden layers. In addition, a large scale study conducted by Saberironaghi indicated that ANN's have greatly assisted in improving prediction capabilities through their flexibility and ability to adapt.

Naïve Bayes (NB)

Naïve Bayes (NB) is a probabilistic classifier based on Bayes’ theorem where you estimate class posterior probability assuming independence of features [3]. In stock market prediction NB has been applied using technical indicators like RSI, MACD, and moving averages, as features. Empirical work predicting market direction by Patel et al. 2015 show that NB can be a useful stand-in baseline due to its speed of experimentation, good performance especially on low-dimensional but well-preprocessed feature spaces), and utility for making real-time decisions. However, financial indicators are all correlated e.g. SMA and EMA across multiple windows, meaning your estimated probabilities are biased leading to lower prediction performance in high-dimensional datasets.
Benefit: Fast, and works well in a real-time context for “quick and dirty” labeling purposes

Drawback: Naïve feature independence assumption leads to poor performance in correlated feature spaces

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a classical statistic classification method that projects high-dimensional to lower-dimensional space maximizing variance between classes and minimizing within-class variance. It's been applied in financial prediction of distinguishing bullish and bearish regimes [24]. The author in Huang et al. (2005) applied LDA on predicting stock direction using some features that ended up being reasonably Gaussian and separable enough for LDA to work well. LDA yields a linear projection into lower-dimensional space that captures variance between the two classes, while minimizing variance relative to each class. [9] [28] Since financial time series data is generally nonlinear in nature, these derivation methods do not hold strictly. And indeed, stock market data is very non-Gaussian in nature, not even clustering heavy tails in shape.

Method	Core Mechanism	Financial Data Handling Capability	Strengths	Limitations	Best Use Case in Finance
Logistic Regression	Linear decision boundary using sigmoid probability function	Handles basic OHLC and limited technical indicators; struggles with nonlinear patterns	Highly interpretable, fast training, robust baseline model	Cannot model nonlinear relationships; poor performance in volatile markets	Baseline comparison, small datasets, interpretable trading rules
Naïve Bayes	Probabilistic classifier assuming feature independence	Works with simple indicator sets; fails with correlated financial features	Computationally efficient, performs well with limited data	Assumption of feature independence unrealistic in financial markets; ignores interactions	Quick benchmarking, low-dimensional feature spaces
LDA	Linear projection maximizing class separability	Effective when financial features are normally distributed and separable	Reduces dimensionality, improves class separation	Assumes Gaussian distribution and equal covariance; weak for nonlinear dynamics	Structured financial datasets with low noise
SVM	Maximizes margin between classes using kernel functions (RBF, polynomial)	Handles high-dimensional technical indicators and nonlinear relationships	Effective in high-dimensional spaces, robust to overfitting	Sensitive to kernel and hyperparameters; scalability issues	Medium-sized datasets with complex feature spaces

Deep Learning Classification

Recurrent Neural Networks & LSTM's

Recurrent Neural Networks (RNN's) & LSTM's are RNN's that have become the preferred method for predicting stock prices, because of their ability to recognize temporal dependencies in Data, and have been extensively researched. For example Phuoc et al illustrated that LSTM's have shown a high level of predictive accuracy (about 93% accuracy) in predicting price directions of stocks, based on the input of technical indicator data [16]. In addition, the results of comparative analysis performed on LSTM's, have demonstrated that LSTM models consistently outperform traditional methods of predicting Time-Series Data (ARIMA) in terms of their ability to understand and capture non-linear temporal patterns. In most recent years hybrid LSTM models have also been studied, the most notable example being that of Latif et al, who developed an LSTM model that incorporated an attention mechanism [11][26]. This model had a higher degree of accuracy than previous LSTM models due to the ability to focus on the most relevant temporal features when making predictions.

Benefits of LSTM:

- LSTM's are excellent at capturing temporal dependency
- LSTM's can model both Long and Short Term Patterns
- LSTM's demonstrate high levels of predictive accuracy

Drawbacks:

- LSTM models require a large amount of Data, thus have a high computational cost
- LSTM models are very susceptible to overfitting, especially in noisy market environments.

Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) Kyunghyun Cho et al., 2014) originated as a computationally less expensive alternative to LSTMs, merging the forget and input gates into a single update gate while also eliminating memory cells altogether. GRU has gained traction in stock market prediction research, used on financial time-series data due to its ability to learn from past observations of prices, returns, and technical indicators. Results by Thomas Fischer and Christopher Krauss (2018) show that GRU has performance roughly on par with LSTMs, but fewer parameters and a shorter time needed to train the models [18]. This makes them especially attractive when working with large datasets in finance. Similar to LSTM with daily stock prices and technical indicators as inputs, GRUs learn tendencies across windows of time in phenomena such as short-term momentum and long-term trends. The gating technique within these architecture gives GRUs the ability to zero in on only relevant stored observations of extreme value and reject the rest with no positive value being provided. This memory outer is particularly valuable given volatility of the stock market. There also have been variations of this idea combining GRUs with attention mechanisms, or in hybrid architectures like CNN-GRU or a mixture of different recurrent approaches with multiple types of layers.

Benefits:

- GRU requires fewer parameters and less training time than LSTMs

Drawbacks:

- Still may not be as expressive as LSTM in capturing long-term dependencies

Attention mechanisms

Attention mechanisms first introduced by Dzmitry Bahdanau and others as early as 2015, opened up a whole new field of the deep learning models. These mechanisms enable models to place different weighting of input features and time steps and place more emphasis on the features and time steps that really matter to the prediction. Attention based models of stock market forecasting are able to recognize and highlight the significant time range or technical indicator, virtually the needle in a hay stack that actually moves the market. The combination of attention mechanisms and recurrent neural networks, e.g. LSTMs or GRUs, is a common choice of attention in the case of financial time-series classification. One of them is Take Yao Qin dual-stage attention-based network of 2017. It selectively attends to the features and the crucial moments, which triggers the enhancement of the prediction accuracy. More recently, in 2021 Bryan Lim and his colleagues introduced the Temporal Fusion Transformer applying attention to make clear forecasts that are understandable across diverse time horizons. There is more to magic of attention in finance than performance. It is regarding the openness of models. Attention models enable us to see what predictors of price change, be it RSI, MACD, volatility measures, etc., or what time periods actually predict where the price will be. The latter form of transparency is quite appropriate, in the era of the growing popularity of explainable AI in finance. One more enormous plus: attention mechanisms address a traditional flaw of deep learning i.e. the same treatment of all inputs. Markets are dynamic and what is important today may change. Attention enables models to vary attentively to the shifting cues instead of being attached to the old patterns. So, the benefit? It makes the interpretation of models easier and allows their adaptation to the changing information and time. On the negative, the extra consideration would mean increased complexity and extra calculations in the model.

Auto encoder-Based Classification

In 2006, Geoffrey Hinton and Ruslan Salakhutdinov proposed unsupervised neural networks that learn to compress data into meaningful and compact representations, and are referred to as autoencoders. With dimensionality reduction, they capture the gist of the input. Autoencoders are frequently used in stock prediction: they are used to extract latent features out of large, unstructured financial data and then the extracted features are given to classification models. The financial data is infamous in terms of the abundance of technical indicators, and quite often within time windows. That generates a mess of redundancy and correlated features and high likelihood of overfitting. Autoencoders come in handy.

Method	Core Mechanism	Financial Data Handling Capability	Strengths	Limitations	Best Use Case in Finance
ANN	Multi-layer feed forward network with nonlinear activation functions	Captures nonlinear relationships among technical indicators	Flexible architecture, models complex feature interactions	Requires large data, prone to overfitting, lacks temporal memory	Nonlinear pattern recognition in structured financial data
RNN	Sequential modeling using recurrent connections across time steps	Captures short-term temporal dependencies in stock prices	Suitable for sequential data, models time dependencies	Suffers from vanishing gradient problem; weak long-term memory	Short-term time-series prediction

LSTM	Memory cell with gating mechanism (input, forget, output gates)	Effectively models long-term dependencies in financial time-series	Strong temporal modeling, handles nonlinearity and sequence patterns	Computationally expensive; requires tuning and large datasets	Long-term stock trend prediction and sequential modeling
GRU	Simplified gating mechanism combining forget and input gates	Captures temporal dependencies with reduced parameter complexity	Faster training, fewer parameters, efficient for real-time systems	Slightly less expressive than LSTM for long-term dependencies	Real-time prediction systems and large datasets
Attention Models	Assigns weights to important features/time steps dynamically	Identifies key indicators and critical time periods in financial data	Improves interpretability, captures dynamic feature importance	High computational complexity; requires careful tuning	Feature importance analysis and hybrid deep models

They reduce the size of the input by squashing it into a smaller dimensional space in which the noise and unnecessary information is removed-the significant signals are preserved. They can subsequently be inputted into classifiers such as SVM, Random Forest, or deep neural networks, and the learned features can be sent to them, which form the foundation of more powerful prediction. Studies indicate that both stacked and denoising autoencoders provide a good improvement to stock market predictions.Autoencoders are able to learn the nonlinear dynamics of features, yielding a meaningful feature space compared to classical algorithms like PCA. And that is not all other researchers integrate autoencoders with LSTM or GRU networks and, in such a manner, the models are able to extract features and process time patterns at the same time. This type of hybrid approach elevates the performance of classification of financial time-series data. A disadvantage of autoencoders is that they are unsupervised. They are not directly goal-oriented to the goal of prediction and so their features learned may not be exactly what the classification task wants. That is an actual drawback in certain instances. The upside? Autoencoders are good at dimensionality reduction, and capture nonlinear relationships between features. The downside? They are not necessarily optimized in their features to classify.

RESEARCH GAPS AND FUTURE SCOPE

Nonetheless, despite all the modernization in the categorization of stock markets, there are some research gaps, which appear to be glaring [6] [22]. Their preoccupation with maximizing predictive accuracy-and omit temporal stability-is found in most models. That results in drops in performance when there is a change in the regime of the market. Another issue: models are models which take technical indicators as independent variables, totally disregarding the web of interactions and dependencies which cause the real financial dynamics [23]. High dimensional, multi-window features have only a tendency of adding redundancy and multicollinearity to the model, which increases the chances of overfitting and reduces their ability to generalize.

Better still, there are loads of studies employing wobbly validation strategies. It then results in data leakage and way too optimistic results that casts the validity of such findings into doubt. Future research has to fill the existing gaps through focusing on feature selection depending on stability awareness, interactivity, mix hybrid deep learning with explainability and serious time-series validation like walk-forward evaluation. These are important steps to take models out of the lab and actually to make them work in the real markets, which are ever-changing and financial.

CONCLUSION

The paper adopted an in-depth overview of the classification procedures in stock market prediction, starting with the classics of Support Vector Machines and ensemble methods, and extending to the latest state-of-the-art deep learning techniques. SVMs and ensemble models are best with structured data, but deep learning models, including LSTM, GRU and attention-based models, are in another league to detect nonlinear trends and time effects in financial data. Nonetheless, the review also indicates that a major issue is that most of the models rely on high scores on historical data, and neglect the problems of non-stationarity, feature instability, overfitting, and sketchy time-series validation. Moreover, the lack of consideration of interactions between the features, and redundancy of high-dimensional data undermine robustness and limits the capacity of these models to perform well in reality. Absolute accuracy should not be sought as a goal of future studies. The field demands fixed, interrelation-sensitive, and strong frameworks that can handle the market dynamics. Those developments are essential in case we wish to bridge the disparity between theoretical victories and actual, dependable application in the whirlwind financial markets.

REFERENCES

1. M. Saberironaghi, J. Ren, and A. Saberironaghi, "Stock market prediction using machine learning and deep learning techniques: A review," AppliedMath, vol. 5, no. 3, 2025.
2. T. Ren, "Stock market forecasting based on machine learning and investor sentiment," Physica A, 2025.
3. J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," Expert Systems with Applications, vol. 42, no. 4, pp. 2162–2172, 2015.
4. G. Sonkavde et al., "Forecasting stock market prices using machine learning," Journal of Risk and Financial Management, vol. 16, no. 3, 2023.
5. C. Y. Lin, "Stock market prediction using artificial intelligence: A systematic review," Heliyon, vol. 10, 2024.
6. M. Nabipour et al., "Predicting stock market trends using machine learning and deep learning algorithms," Neural Computing and Applications, vol. 32, 2020.
7. W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using neural networks," Information Sciences, vol. 2025.
8. Y. Karulkar, "Evaluating machine learning models for stock market prediction using sentiment analysis," SAGE Journals, 2025.
9. A. Rohan et al., "Artificial intelligence in financial market prediction: A review," Frontiers in Artificial Intelligence, 2025.
10. T. Chaudhary, "Machine learning and time-series approaches for stock market prediction," Procedia Computer Science, 2025.
11. J. Chen et al., "Machine learning techniques for stock price prediction and signal identification," Engineering Applications of Artificial Intelligence, vol. 2023.
12. W. Sun et al., "Deep learning models for stock prediction and financial forecasting," Scientific Reports, Nature, 2025.
13. V. Kumar, "A review of supervised machine learning methods for stock prediction," IRJET, vol. 12, no. 4, 2025.
14. O. Bustos et al., "Machine learning, stock market forecasting, and market efficiency," Soft Computing, 2025.
15. L. Del Nero et al., "Machine learning models to predict stock market spillovers using RNN and GRU," Finance Research Letters, 2025.
16. A. Subasi et al., "Stock market prediction using machine learning classifiers," Procedia Computer Science, 2021.
17. S. Vadlamudi, "Stock market prediction using machine learning: A systematic literature review," 2017.
18. Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, "Deep reinforcement learning for financial trading," IEEE Computational Intelligence Magazine, vol. 11, no. 4, pp. 40–56, 2016.
19. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria," IEEE TPAMI, vol. 27, no. 8, pp. 1226–1238, 2005.
20. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.
21. C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
22. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
23. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
24. K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," EMNLP, 2014.
25. Y. Qin et al., "A dual-stage attention-based recurrent neural network for time series prediction," IJCAI, 2017.
26. B. Lim et al., "Temporal fusion transformers for interpretable multi-horizon time series forecasting," International Journal of Forecasting, vol. 37, no. 4, 2021.
27. G. Hinton and R. Salakhutdinov, "Reducing dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.
28. T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," European Journal of Operational Research, vol. 270, 2018.
29. B. Mandelbrot, "The variation of certain speculative prices," Journal of Business, vol. 36, 1963.
30. E. F. Fama, "Efficient capital markets: A review of theory and empirical work," Journal of Finance, vol. 25, no. 2, pp. 383–417, 1970.J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock market index using fusion of machine learning techniques," Expert Systems with Applications, vol. 42, no. 4, pp. 2162–2172, Mar. 2015.