International Journal of Advanced Research in Commerce, Management & Social Science (IJARCMSS)
(An International Quarterly Multidisciplinary Open Access Double Blind Peer Reviewed Refereed Journal)
ISSN :2581-7930 (Online), Impact Factor (Cosmos): 7.270 & PIF (I2OR): 5.010, Volume 09, No. 04(I), October-December, 2026, pp 175-180

Big Data Analytics on Stock Price Prediction: A Comparative Study of Machine Learning and Deep Learning Models (2018-2026)

Dr. Deepika Chaplot[1] | Riya Chatterjee[2*]

[1]Associate Professor, Department of Commerce, Pacific Academy of Higher Education and Research University.
[2]Student, Department of Commerce, Pacific Academy of Higher Education and Research University.

[*]Corresponding Author: riya.chatterjee0412@gmail.com

Citation: *Chaplot, D. & Chatterjee, R. (2026). Big Data Analytics on Stock Price Prediction: A Comparative Study of Machine Learning and Deep Learning Models (2018-2026). International Journal of Advanced Research in Commerce, Management &Amp; Social Science, 09(04(I)), 175–180. https://doi.org/10.62823/ijarcmss/9.4(I).9708*

Abstract

The issue of predicting stock prices is no secret. This review looks at papers from 2018-2026 that look at various models both from a traditional ML perspective (random forests, gradient boosting, support vectors) and from a deep learning perspective (LSTM, graph networks, transformers) and how they can be applied to big data analytics. Overall deep models offer superiority on richer/higher frequency data whereas tree based models offer continual competitiveness on tabular data. However, evaluation can be very important and issues such as leakage and not considering transaction cost can occur. Section 6 provides some key findings. Section 7 details some open questions. Finally, there is a brief discussion of what is going on at a broader level within India regarding data analytics, AI adoption and financial institutions.

Keywords: *Stock Price Prediction, Big Data Analytics, Machine Learning, Deep Learning, LSTM, Transformers, Financial Forecasting.*

Introduction

Half the finance students will say no due to the efficient market hypothesis. The other half will open their Jupyter notebook and hope no one sees. Both answers are correct. The efficient market theory suggests that all available information will be reflected in prices. However, in reality there will always be some lag. If we have large amounts of data and computing power we can exploit minor regular inefficiencies.

Since 2018 we have had an abundance of data. We have large amounts of order book data, news feeds, earnings call transcripts, text from social media and alternative data (satellite imagery, card data). Big data analytics requires a pipeline capable of handling high volumes, high velocity and high variety of data. Models must be capable of using all the data.

This paper aims to understand how ML and DL approaches are used to predict stock prices within literature in the period 2018-2026 and whether deep models add any extra value. This paper is a narrative review so is not an official meta-analysis. The models used within this period varied greatly in terms of market type, time horizon, features and what was measured so we cannot determine which model is the best. If attempting please be cautious!

Literature Selection

Peer reviewed journals and high quality conference papers from 2018 onwards. The papers must have made at least 2 families of models to predict stocks/indices or provided a review. Some papers were included to give background or practical knowledge. Some papers were included from India

Copyright © 2026 by Author's and Licensed by Inspira. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work properly cited.

regarding management and financial institutions. These did not focus on stock prediction but focused on data analytics, AI adoption and the surrounding institutions. These have been detailed in section 8 as it would be misleading to consider them evidence.

Machine Learning

- **Tabular Data**

Probably the most influential paper from the ML side is Gu, Kelly and Xiu (2020). This paper tested a variety of models (penalised regression, random forests, boosted trees, neural networks) to predict cross-sectional returns. This found that non-linear models were best with tree based and shallow neural networks being best. This was because they could account for variable interactions. Interestingly the best neural network was quite shallow!

Outside of finance this has been proven! Grinsztajn, Oyallon and Varoquaux (2022) found that tree based models continue to outperform deep networks on many tabular data sets. This is because the data is irregular with many uninformative features. Neural networks suffer here. Although this does not look at stock prediction many data sets will contain tabular data such as valuations ratios, momentum and volume.

- **ML Reviews**

Nti, Adekoya and Weyori (2020) created a review looking at fundamental and technical analysis across a number of years. They found that technical indicators were dominant with SVMs and neural networks being the most common. Kumbure, Lohrmann, Luukka and Porras (2022) created an up to date review of techniques and data sets. They concluded that whilst there is a large variety of models there is less standardisation in data, benchmarking and validation. Sometimes it can be unclear whether the 5% increase is due to the model, features or just luck in the test period.

- **Overfitting**

Not an empirical paper but worth mentioning. Lopez de Prado (2018) argues that due to the low signal to noise ratio in financial data, if enough strategies are tried you will find one that looks extremely successful. This means many accuracy measures are not reliable. Especially ones claiming very high directional accuracy (70-90%) on a single stock.

Deep Learning

- **Recurrent Neural Networks**

Fischer and Krauss (2018) used LSTMs to predict return direction for constituent stocks in the S&P 500. They compared this to random forests, a deep feed forward network and logistic regression. They found the LSTM model performed best however this reduced throughout the years showing that the market was adapting. This was one of the cleanest comparisons with a large universe and consistent trading rules.

Sezer, Gudelek and Ozbayoglu (2020) created a survey of deep learning methods for financial time series forecasting between 2005-2019. They found LSTMs and CNNs dominated. Many papers did not use much data with poor baselines. Jiang (2021) created a similar review for stock prediction using deep learning. They found whilst architectures have improved quickly, evaluation was not consistent.

- **Graph Networks**

Stocks rarely move independently. There are many links between stocks. Feng et al (2019) created a temporal relational ranking model using sequential methods and graph relationships. This outperformed models assuming stocks moved independently. Kim et al (2019) also used a hierarchical graph attention network to predict stock movement. This is where DL really adds value as it is difficult to create relationships manually with a tree model.

- **Transformers**

Transformers have taken a while to enter finance. Lim, Arık, Loeff and Pfister (2021) created the Temporal Fusion Transformer for multi-horizon forecasting. This works well and can be interpreted with a mix of static and time varying inputs. However Zeng, Chen, Zhang and Xu (2023) showed very simple linear models can outperform or comparable to several transformer based forecasters on standard long-horizon time series benchmarks. This paper is not related to equities and the authors were not stating transformers were not useful. However it is a good reminder that newer architectures do not always produce good results in forecasting, especially with noisy financial series.

A Side-by-Side View

Table 1: Classical ML Versus Deep Learning: A Rule-of-Thumb Comparison

Dimension	Classical ML (Trees, SVM, Regularized Regression)	Deep Learning (LSTM, GNN, Transformer)
Best-suited data	Tabular, engineered features, moderate samples	High-frequency, text, relational, long sequences
Training cost	Low to moderate	Moderate to very high
Interpretability	Relatively good (feature importance)	Weaker, though attention helps somewhat
Sensitivity to noise	More robust on noisy tabular features	Prone to overfitting without heavy regularization
Typical evidence	Strong for cross-sectional return prediction (Gu et al., 2020)	Strong when relations or text are modeled (Feng et al., 2019; Fischer & Krauss, 2018)
Main risk	Plateau in accuracy	Spurious backtest performance

Source: Author's synthesis of the reviewed literature.

I would not read this table as a verdict. It is closer to a rule of thumb. If your data is a clean tabular panel and your team is small, boosted trees are a defensible starting point. If you are mixing news text, price history and inter-firm links, the deep route is probably worth the pain.

Essential Data from the Reviewed Studies

Narrative comparisons are easier to trust when the numbers sit next to them. This section pulls together the study designs and the headline figures reported in the two most-cited empirical comparisons. A caution first: these are different markets, horizons and metrics, so the numbers should be read within each study, not ranked against each other.

- **Study Overview**

Table 2: Overview of the Reviewed Studies

Study	Year	Type	Data / Market	Main Takeaway
Fischer & Krauss	2018	Empirical	S&P 500 constituents, daily, 1992-2015	LSTM beat RF, DNN and logistic regression before costs; edge faded later
López de Prado	2018	Book / method	General financial ML practice	Backtest overfitting and weak validation inflate results
Feng et al.	2019	Empirical	US stocks (NASDAQ, NYSE) with relation data	Adding stock relations to sequential models improves ranking
Kim et al.	2019	Empirical	Stock movement with graph-based relations	Hierarchical graph attention helps movement prediction
Gu, Kelly & Xiu	2020	Empirical	Large US stock panel, monthly, decades of data	Nonlinear models (trees, shallow NNs) beat linear ones
Nti et al.	2020	Review	Fundamental and technical analysis studies	Technical indicators, SVMs and neural nets dominate
Sezer et al.	2020	Review	Deep learning forecasting studies, 2005-2019	LSTM and CNN dominate; weak baselines are common
Lim et al.	2021	Method	Multi-horizon time series benchmarks	Temporal Fusion Transformer: accurate and interpretable
Jiang	2021	Review	Deep learning for stock prediction	Fast architectural progress, inconsistent evaluation
Kumbure et al.	2022	Review	ML techniques and data sources	Little standardization of data, benchmarks, validation
Grinsztajn et al.	2022	Benchmark	Many tabular datasets (non-financial)	Trees often outperform deep networks on tabular data
Zeng et al.	2023	Benchmark	Long-horizon time series (non-financial)	Simple linear models rival transformer forecasters

Source: Compiled from the reviewed papers. Markets and horizons differ across studies.

- **Reported Results: Gu, Kelly and Xiu (2020)**

Gu et al. (2020) scored each method on monthly out-of-sample R-squared for individual stock returns. The absolute numbers look tiny, and that is the point: monthly stock returns are mostly noise, so even a fraction of a percent is economically meaningful when applied across thousands of stocks. The ordering is what matters. Linear methods sit at the bottom, tree ensembles in the middle, and shallow neural networks at the top, with deeper networks (four and five layers) not improving on three.

Table 3: Approximate monthly out-of-sample R-squared (%) by model, Gu et al. (2020)

Model	R^2 (%)	Model	R^2 (%)
OLS (3 predictors)	0.16	Random forest	0.33
Elastic net	0.11	Gradient boosted trees	0.34
Generalized linear model	0.19	Neural net, 1 layer	0.33
Principal components regression	0.26	Neural net, 2 layers	0.39
Partial least squares	0.27	Neural net, 3 layers (best)	0.40
		Neural net, 4 layers	0.39
		Neural net, 5 layers	0.36

Note: Values are approximate, reproduced from memory of the published tables; verify against Gu et al. (2020) before submission.

- **Reported Results: Fischer and Krauss (2018)**

Fischer and Krauss (2018) traded a long-short portfolio built from each model's daily predictions on S&P 500 constituents. The LSTM earned the highest mean daily return before transaction costs, with the random forest close behind. Two lessons follow. First, the gap between the LSTM and random forest is small, which sits comfortably with the argument that tree methods remain strong competitors. Second, the authors reported that profitability declined in later years, which is consistent with growing market efficiency or crowding.

Table 4: Approximate Mean daily Returns before Transaction Costs, S&P 500, 1992-2015

Model	Mean Daily Return (%)
LSTM	0.46
Random forest	0.43
Deep feedforward network	0.32
Logistic regression	0.26

Note: Approximate figures for the top-k portfolio as recalled from the paper; verify against Fischer & Krauss (2018). Returns are before transaction costs.

- **Data Sources and Evaluation Metrics**

The studies above draw on a handful of recurring data types and are judged by a handful of recurring metrics. Tables 5 and 6 summarize both, since mismatches here explain much of the disagreement between papers.

Table 5: Common Data Sources in Big Data Stock Prediction

Data Type	Examples	Typical Challenge
Price and volume	Daily or intraday OHLCV, order-book snapshots	Low signal-to-noise; non-stationarity
Technical indicators	Moving averages, RSI, MACD, volatility	Highly correlated features
Fundamentals	Earnings, valuation ratios, balance sheet items	Low frequency; reporting delays
Text and sentiment	News, earnings call transcripts, social media	Noise; signal decays fast
Relational data	Sector, supply chain, ownership links	Hard to build and keep current
Alternative data	Satellite imagery, web traffic, card spending	Cost; access; coverage bias

Source: Author's synthesis of Kumbure et al. (2022), Nti et al. (2020) and Sezer et al. (2020).

Table 6: Common Evaluation Metrics and their Limits

Metric	What it Measures	Main Limitation
RMSE / MAE	Size of price or return error	Small error can still lose money
Out-of-sample R-squared	Share of return variance explained	Tiny values are normal; easy to misread
Directional accuracy	Share of up/down calls right	Ignores size of moves and costs
Sharpe ratio	Return per unit of risk	Sensitive to window and assumptions
Net returns after costs	Profit after fees, slippage, impact	Rarely reported, rarely realistic

Source: Author's synthesis.

Big Data Considerations

Most of the literature relating to big data focuses on the infrastructure. Issues such as handling tick data, text cleaning and temporal alignment can become quite dull. An issue with leakage can occur during this process. If, by mistake, any information after the prediction time stamp is used in the features your model will appear as a genius. This is highlighted by both Kumbure et al. (2022) and Lopez de Prado (2018). Maybe you should also consider whether more data is actually better. Often with text and alternative data sources, the signal will be lost quickly as more people start using it. This is true for Fischer and Krauss' (2018) strategy. However this can be seen quite loosely as there are many reasons why this could happen. 8. Related Work in Indian Management and Finance None of the following papers give predictions for stock price. They have been included based on the topic. Think of this as background information. Chaplot et al. (2023) looked into Customer Relationship Management in the age of Data Analytics. This shows that within Indian companies analytics are being used more and more as a decision making tool. Exactly what is needed for our stock prediction pipeline. Nagda and Mehta (2026) gave an example of how HDFC are using their EVA chatbot as a potential replacement for human customer support. This again shows that AI is becoming more popular in Indian finance. If banks are comfortable with using AI to manage their customers, then how trusting should we be of algorithms in investment decisions? Joon (2026) looked into how startups and traditional companies are embracing ESG and AI strategies to stay ahead in competition. This shows how the adoption of strategies is not balanced. If an organisation does not have clean data and people managing it then a complex DL prediction model won't help. Abid (2026) looked into how increased access to finance for SME's is driving Aatmnirbhar Bharat and Kotak and Paliwal (n.d.) looked into how Indian banks are implementing strategies to recover from NPA. Both highlight that Indian markets have institutional characteristics that aren't covered in US focused papers (Gu et al., 2020 and Fischer and Krauss, 2018). A model trained on S&P500 stocks may not perform well on data from NSE/BSE. This seems to be an area that hasn't been heavily explored.

Critiques and Improvements Some points raised whilst reading the papers include: Many of the papers focus on just accuracy/RMSE. Few of them mention incorporating transaction costs, slippage and market impact into their return calculations. A model with 55% accuracy may not be profitable. Zeng et al. (2023) show that the simplest baseline should be treated seriously! It would have been nice to see more structure within the finance papers. Comparing a tuned LSTM with an untuned linear model isn't a fair comparison. Many of the models have been trained until 2018/2021 meaning they have experienced the pandemic crash, stimulus rally and rate hikes. Few papers have mentioned stress testing their models. There is a big bias in favour of US data throughout the literature. Whilst there are some papers which are looking at Indian data, Brazilian data etc. there is not much and the microstructure will be different. A deeper point is if the model was actually effective and became popular, would it still be effective? Probably not for very long. This means that the question of which model is better isn't static as it would be with something like image classification. 10. Conclusion There is no one model family that outperforms others. Gradient boosted trees and shallow networks perform well for cross-sectional tabular data (Gu et al., 2020; Grinsztajn et al., 2022), but relational, text or lengthy sequence data shows clearer benefits of using deep models (Feng et al., 2019; Fischer & Krauss, 2018; Lim et al., 2021). Transformers look like they could be a good option but have not been proven to provide a significant improvement on noisy financial series (Zeng et al., 2023). Overall the main problem within this field is validating the results. Many reviews have highlighted this (Jiang, 2021; Kumbure et al., 2022; Nti et al., 2020; Sezer et al., 2020). The most useful improvements would be having standard benchmarks, cost aware validation and looking into more Indian and other emerging markets.

References

1. *Abid, M. (2026). Empowering small and medium enterprises (SMEs) through access to finance: A key driver of Aatmnirbhar Bharat. Prayatna: A Journal of Research, 1(1), 46-51.*
2. *Chaplot, D., Ranawat, P., Yadav, A., & Soni, K. (2023). Customer relationship management (CRM) in the era of data analytics. European Chemical Bulletin, 12(4), 18392-18406.*
3. *Feng, F., Chen, H., He, X., Ding, J., Sun, M., & Chua, T.-S. (2019). Temporal relational ranking for stock prediction. ACM Transactions on Information Systems, 37(2), Article 27.*
4. *Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654-669.*

5. *Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35 (NeurIPS Datasets and Benchmarks Track).*

6. *Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273.*

7. *Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. Expert Systems with Applications, 184, Article 115537.*

8. *Joon, A. A. (2026, April 24-25). Startups vs traditional firms: ESG & AI adoption for competitive growth [Conference paper]. 17th International Conference on ESG and AI for Sustainability & Inclusivity, Pacific Academy of Higher Education & Research University, Udaipur, India.*

9. *Kim, R., So, C. H., Jeong, M., Lee, S., Kim, J., & Kang, J. (2019). HATS: A hierarchical graph attention network for stock movement prediction. arXiv preprint arXiv:1908.07999.*

10. *Kotak, S., & Paliwal, R. (n.d.). A study of NPA recovering measures in Indian banks. Unnant: The Business Journal, 10(1), 1-6.*

11. *Kumbure, M. M., Lohrmann, C., Luukka, P., & Porras, J. (2022). Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems with Applications, 197, Article 116659.*

12. *Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748-1764.*

13. *López de Prado, M. (2018). Advances in financial machine learning. Wiley.*

14. *Nagda, S., & Mehta, A. (2026). HDFC EVA chatbot: A case study on its potential to replace human customer support. International Journal of Management, Economics and Commerce, 3(1), 101-111.*

15. *Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A systematic review of fundamental and technical analysis of stock market predictions. Artificial Intelligence Review, 53(4), 3007-3057.*

16. *Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005-2019. Applied Soft Computing, 90, Article 106181.*

17. *Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2023). Are transformers effective for time series forecasting? Proceedings of the AAAI Conference on Artificial Intelligence, 37(9), 11121-11128.*

❑❑❑