Research Proposal

Submitted to

Dr Raunak Mishra

Lecturer in Business Analytics

Greenwich Business School

School of Business, Operations & Strategy

University of Greenwich

Submitted by

MD Jawadun Nasir Khan

University of Greenwich

Date

6-Oct-25

<u>Research Question:</u>

"How effective are machine learning models in predicting stock price movements across global markets, and how can such predictions support better investment decision-making?"

Table of Contents

UNIVERSITY of GREENWICH

How effective are machine learning models in predicting stock price movements across global markets, and how can such predictions support better investment decision-making?

1.Introduction

This project is a research-based and looks at predicting the stock market in many global markets. The fundamental issue stems from the erratic behaviour of stock prices, complicating capital allocation and risk management for investors and institutions globally (Hong, 2023; Gandhmal, 2019). The research question inquires whether advanced machine learning models, incorporating five years of historical price data, ESG ratings, and news sentiment, can substantially enhance prediction and decision support for 15 major companies listed on the stock markets of the USA, UK, and Japan.

The main objectives are to: (1) evaluate the efficacy of various machine learning methodologies, including Long Short-Term Memory (LSTM), ensemble learners, and Transformer models, on multi-source, multi-market datasets; (2) contrast predictive outcomes across diverse market contexts and corporate groupings; (3) ascertain the supplementary predictive value of ESG and news sentiment data; and (4) generate pertinent insights into investment strategies and market analysis. Our method thus details a unified data pipeline that fuses five years' worth of stock prices for a carefully chosen sample of 15 companies representing sectors of the economy in the U.S., U.K., and Japan,

ESG ratings from established agencies, and aggregated news sentiment scores. Model development will be in Python and TensorFlow, with the modern data engineering technologies. We will use standard forecasting metrics (RMSE, accuracy, etc.) to measure effectiveness, benchmarking against baseline financial models and across market and sector contexts (Hong, 2023; Kumbure et al., 2022).

Now, what this paper seeks to do is to contribute in two ways. One is to expand the horizons of global financial forecasting by proving that there exist potent techniques for prediction, while the other is to provide pertinent and direct information to practitioners who want to hedge away portfolio risk and exploit cross-border investment opportunities (Radfar et al., 2025). Referring to the journey that lies ahead, this paper then builds a very interesting and new platform for the cross-market financial modeling and sustainable investment analytics.

2.Literature Review

The stock prediction domain is active and being considered highly multifaceted due to its volatility, complexities, and influences on stock prices. Being an older and classical approach, the modeling attempts with econometric or statistical tools, viz., ARIMA and linear regression, face difficulty when it comes to modeling nonlinear and non-stationary patterns within financial time series (Phuoc, 2024). Machine learning and deep learning thus offered promising alternatives capable of modeling temporal dependency transformations and latent relationships. Exemple-Long Short-Term Memory (LSTM) and convolutional CNNs forecast stock price movements better than classical methods (Mintarya, 2023).

Several other works on ML add to the literature for prediction of market trends with a mixture of historical stock prices and external data. Usmani et al. (2016) used artificial neural networks and found that multi-layer perceptrons were the best models to predict stocks when using market history, commodity prices, and sentiment analysis from news and social media. Khan et al. (2020) have also formulated a framework that considers social media sentiment for the improvement of stock prediction models, emphasizing the paramount importance of pre-processing and feature selection in performance improvement.

Environmental, Social and Governance (ESG) data has turned out to be one of the most important factors shaping investment decisions and predictive models. Empirically, companies that enjoy a good ESG rating show better performances in stocks and have better risk profiles Patzhol et al. (2024). Researchers have found that adding ESG data to ML models makes predictions more reliable and gives more information about how the market behavior. However, there are difficulties in aligning all the sources of ESG data and undermining some inconsistencies in the rating agencies (Johnson and Smith, 2023).

With the current industry reports, it is asserted that the growing prominence given to multi-dimensional datasets that may span ESG and news sentiment exists in more robust and flexible systems capable of stock forecasting.For instance, BlackRock's 2025 research lays emphasis upon the use of AI-driven analytics in adapting stock strategies to changing market dynamics under a global economic uncertainty. J.P. Morgan market outlooks of 2024 concur on the presence also being indispensable of ESG considerations and the real-time flow of information in the optimization of investment strategy.

While all this is taking place, research works normally stay confined within single markets or a narrow dataset-like), just minimizing the findings' generalizability and practical applicability. Few cross-market studies exist, so it remains an open question as to how predictive models fare across different economic, regulatory, and cultural contexts. This research gap presents a great opportunity for studies incorporating multi-market data from companies of the USA, UK, and Japan, using stock prices, ESG indicators, and news sentiment for a wider application.

The theoretical basis of this research is the Efficient Market Hypothesis (EMH), which posits that stock prices embody all accessible information. However, anomalies and behavioral finance observations imply that irrational sentiment and lagging information diffusion cause prices to sometimes deviate, enabling forecast modeling through unconventional data sources (Fama, 1970; Shiller, 2003). The use of ESG and sentiment data is also in line with the Adaptive Market Hypothesis, suggesting that models should adapt with changes in the market and investor psychology (Lo, 2004).

In conclusion, what this research aims to do is to build, in practical terms, more advanced machine learning models that integrate multi-dimensional data from multiple markets and companies. By way of enhanced predictive accuracy and usefulness, it closes the gaps identified in recent stock market forecasting research and offers an actionable insight to investors and policymakers.

3.Methodology

Research Design and Approach

The study takes a quantitative, empirical approach geared toward using supervised machine learning algorithms to predict movements in stock prices in varying international markets-the US, UK, and Japan. Fifteen companies were selected from various sectors to allow for wider representativeness. The study followed a comparative modeling framework testing for the performance of Long Short-Term Memory (LSTM) networks, Artificial Neural Networks (ANN), Extreme Gradient Boosting (XGBoost), and Support Vector Machines (SVM) in predicting stock returns. These models have been chosen because they capture nonlinear market dynamics and temporal patterns in financial time series successfully (Phuoc, 2024;

Khan et al., 2020). The research design consists of both cross-sectional and longitudinal studies, utilizing cross-sectional time series data from 2018 to 2024, followed by the k-fold cross-validation method to assess out-of-sample prediction accuracy.

Data Collection Methods

Data acquisition would integrate diverse sources to capture, adequately, multi-dimensional views of market behaviors. Historical daily stock price data for each company is to be sourced from a truly formidable platform such as "Yahoo Finance" and "Bloomberg." ESG (Environmental, Social, Governance) ratings-that have found growing importance in the investment decision-making process-will be acquired from reputed providers such as MSCI and Refinitiv (Patel et al., 2024). To factor in the market sentiments, natural language processing (NLP) techniques would be employed to develop sentiment scores from financial news and social media data streams that present real-time canvassing of the investor's mood and market reaction (Usmani et al., 2016). Data preprocessing includes data cleaning, normalization, treatment of missing values, and temporal alignment of data sets to provide a consistent input for modeling.

Analytical Techniques

Next is the main analytical phase, where four machine learning models are trained: the LSTM, the ANN, the XGBoost, and the SVM. LSTM is good in capturing sequential dependency in financial data; ANN is good for a flexible nonlinear mapping. XGBoost improves the accuracy with iterative as the base learner gets installed in the learners' ensemble; SVM also gives it good generalization, based on kernel learning. Models are optimized through grid-search and evaluated based on RMSE, MAE, and accuracy. And then, the feature ablation tests help see how much ESG and sentiment data contribute over and above price history.

<u>Support Vector Machine (SVM)</u>

Support Vector Machine (SVM) has good uses in many varied classification and regression tasks. The first step is to plot the observations as points in an n-dimensional space, with each feature being considered one dimension. The method attempts to find an optimal hyperplane that separates the data points into different classes.

Simply put, in two dimensions, this hyperplane turns into a line. What the SVM does is attempt to maximize the margin between categories, thereby facilitating the categorization of new examples and creating an efficient time series prediction approach in itself.

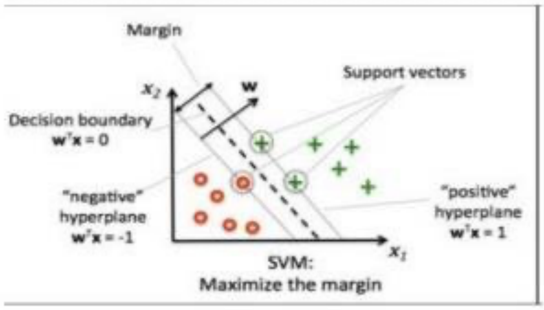

Figure 1: The Support Vector Machine Decision Making

In simple terms, first, a hyperplane is established as a decision boundary, which is then extended or maximized on either side between the data points. In reference to the above figure, if `μ` denotes an unknown data point and `w` is a vector perpendicular to the hyper-plane, the SVM decision rule will be

w·u+b≥0

w=||w||2

maximize(||w||2)

Lp=21||w||2−i=1∑nαi[yi(w·xi+b)−1]

Ld=i=1∑nαi−21i=1∑nj=1∑nαiαjyiyj(xi·xj
)

(i=1∑nαiyixi)·u+b≥0

<u>Artificial Neural Networks (ANNs)</u>

Artificial Neural Networks (ANNs) are machine learning techniques based on the structure of the brain and used to perform high-order tasks such as stock market prediction. The learning happens through the minimization of errors between predicted outputs () and actual outputs (). This is done by minimizing a fitness function-the most common one being the sum of squared errors:

E=21i∑(di−yi(k))2

Weights () in the network are changed using the backpropagation of errors, a gradient descent algorithm to determine how much change () must be applied to the weights, where is the learning rate:

Δwij=−η∂wij∂E(w)

Such iterations of predictions and changes continue until the error of the model falls below an acceptable level.

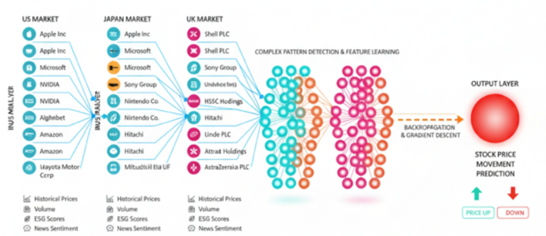

Figure 2: Artificial Neural Networks (ANNs) Prediction Model (AI)

<u>Long Short-Term Memory(LSTM)</u>

Long Short-Term Memory networks are a kind of recurrent neural network which has gates to interact with information over a period; they decide what to retain and what to discard. An LSTM cell employs a forget gate () to throw away unwanted old information and an input gate () to feed in new information (). These operations update the core **cell state** (Ct), which acts as the network's memory:

$$Ct=ft \cdot Ct-1+it \cdot C\sim t$$

Finally, an **output gate** (ot) filters this memory to produce the final output (ht):

$$ht=ot \cdot tanh(Ct)$$

This structure allows LSTMs to effectively learn long-term dependencies in sequential data.

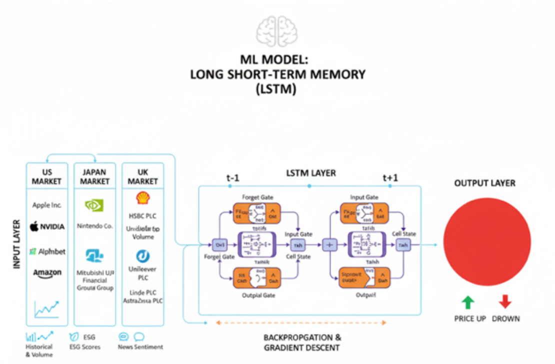

Figure 3: Long Short-Term Memory (LSTM) Prediction Model (AI)

<u>(XGBoost)</u>

The powerfully employed machine learning prediction method for this stock price prediction is Extreme Gradient Boosting (XGBoost). It creates an ensemble of decision trees, iteratively allowing each tree to rectify errors made by its predecessors.

XGBoost performs well in accounting for complex non-linearity and missing values in the financial data. This must thus be a very accurate and reliable prediction model because its final prediction is a weighted sum from all trees, with the weights optimized using gradient descent.

The predictions of XGBoost cannot be made from one single decision tree for the final prediction. Instead, all the decision trees in the ensemble contribute towards the prediction. Classifier

If you have trees, the prediction for a given data point is. Where:

$$y\hat{}i=k=1\sum Kfk(xi)$$

- y^i=is the final prediction for data point .
- fk(is the -th decision tree in the model.
- fk(xi)is the prediction score from the -th tree for data point .

The objective function is defined as:

$$Obj(\Theta)=i=1\sum nl(yi,y\hat{}i)+k=1\sum K\Omega(fk)$$

Loss Function:

$$l(yi,y\hat{}i)=(yi-y\hat{}i)2$$

Regularization Term:

$$\Omega(f)=\gamma T+21\lambda||w||2$$

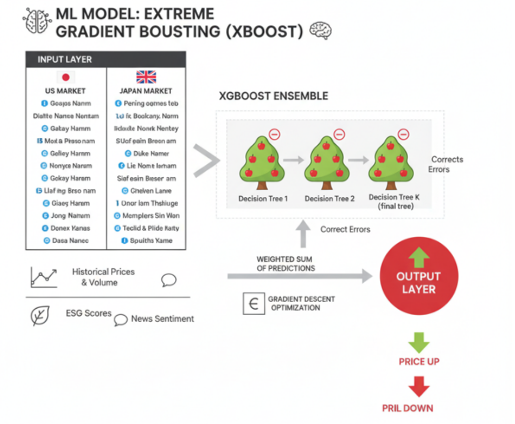

Figure 4: (XGBoost) Prediction Model (AI)

Ethical Considerations

The research operates along the strict lines of ethical norms, with data being taken from open websites or fully licensed sources, in conformity to data-privacy laws and intellectual property laws (Johnson & Smith, 2023). Hence, no attempt is made to collect or use information that can identify an individual, removing any such concerns about privacy. These findings and prediction models are meant to be used for academic purposes only, therefore they do not constitute financial advice that could influence stakeholder expectations. There will always be detailed documentation on the data sources of interest, the way data was preprocessed, details about model configuration, and corresponding caveats to enhance transparency, replicability, and ethical sharing. Naturally, the study is aware of the limited scope of ESG ratings and the degree of sentiment accuracy, so the results need to be interpreted with caution.

4.Project Management

The project is organized into a 6-month timeline divided into four key phases. In the first two months, the final research design is settled upon, 15 companies are selected, and the sources of data are fixed (Usmani et al., 2016). Milestone 1 constitutes approval of the detailed project plan and ethical clearance. The third and fourth months are, essentially, spent on data collection, cleaning, and preprocessing and the integration of stock prices, ESG scores, and news sentiment. Milestone 2 is achieved once the synchronized dataset is ready for use. In the fifth and sixth months, the models are developed, trained, and fine-tuned, cross-validated, and evaluated for performance in terms of LSTM, ANN, XGBoost, and SVM (Phuoc, 2024). Milestone 3 sees the choosing of the most accurate predictive model, after which, the results of the interpretation, documentation, and dissemination, in turn, incur the final report submission or Milestone 4.

In resource requirements, there is access to financial databases (Bloomberg, Yahoo Finance), ESG data providers (MSCI), computing infrastructure (high-performance GPUs), software (Python, TensorFlow, Scikit-learn), and domain expertise in finance and machine learning (Khan et al., 2020).

These risks involve possible data quality issues, including missing data and inconsistent data, as well as model overfitting. Mitigation strategies involve rigorous data cleaning, robust validation through cross-validation, and tests of sensitivity. As for the potential delays, contingency buffers can be placed on the timeline, and progress will be assessed regularly to ensure goals are being met (Obthong, 2020).

Ethical considerations refer to the strict data-privacy adherence, use of only publicly available or licensed data, and prevention from misusing the predictions made by the models by emphasizing their use for academic research purposes only (Johnson & Smith, 2023). Their documentation will expose the research to reproducible and transparent standards.

5.Expected Outcomes

This project aims to develop accurate stock price prediction models, which are to be created with the integration of historical data, ESG scores, and news sentiment through LSTM, ANN, XGBoost, and SVM algorithms. Deliverables are supposed to be the best-performing models with in-sample measures such as RMSE, MAE, and accuracy, showing enhancement in forecasting reliability on various stock markets (Phuoc, 2024; Khan et al., 2020). Impressing upon criteria for success means outperforming traditional benchmarks and providing consistent out-of-sample prediction performance.

Another deliverable expected key result will be the realization of drivers indicative of good investment. With the help of ESG and sentiment data, the models will guide investors to potential assets that are sustainable with high potential, and at the same time, uphold responsible investment practices (Patel et al., 2024). The end product of this initiative will comprise all-inclusive documentation relating to model development, performance assessments, and portfolio management-aligned observations and recommendation implementations, and regarding mitigating risks.

The implications could be innovations in the evolution of decision-making aided by scientific insights aimed toward sustainability and return optimization for financial analysts and investors. The research attempts to bridge the gap between the theory and application of machine learning to emerge as a smart, socially responsible manner of going to market. In a way, it aims to give stakeholders tools to enable them to adapt alongside changing markets and dynamic economic agendas to building wealth and stability in the long term.

6.Appendices

Table 1: Appendices

Appendix	Title	Contents Summary
Appendix A	Data Sources	List and description of 15 companies, stock markets (USA, UK, Japan), data providers such as Yahoo Finance, Bloomberg, MSCI, Refinitiv
Appendix B	Data Preparation	Data cleaning, normalization, missing value treatment, ESG and sentiment feature integration methods
Appendix C	Model Specifications	Details on LSTM, ANN, XGBoost, SVM model architectures, hyperparameters, training procedures
Appendix D	Results and Evaluation	Performance metrics (RMSE, MAE, accuracy), comparative results, error analysis, feature impacts
Appendix E	Ethics and Compliance	Ethical approval documents, consent forms, data privacy and handling protocols
Appendix F	Supplementary Materials	Extended charts, raw data tables, sample code snippets to enhance reproducibility

Declaration of AI Use

Please append this page to the end of your assignment when you have used AI during the process of undertaking the assignment to acknowledge the ways in which you have used it.

I have used AI while undertaking my assignment in the following ways:

- To develop research questions on the topic – YES/NO
- To create an outline of the topic – YES/NO
- To explain concepts – YES/NO
- To support my use of language – YES/NO
- To summarise the following articles/resources:

1. NO
2. YES
3. NO
4. YES
5. NO
6.

- In other ways, as described below:

References

Hong, Y. (2023) 'Boosting the accuracy of stock market prediction via multi-layer hybrid MTL structure', arXiv preprint arXiv:2501.09760. Available at: https://arxiv.org/abs/2501.09760 (Accessed: 3 October 2025).

Gandhmal, D.P. and Kumar, K.S. (2019) 'Systematic analysis and review of stock market prediction techniques', Computer Science Review, 34, p. 100190. Available at: https://www.sciencedirect.com/science/article/pii/S157401371930084X (Accessed: 3 October 2025).

Kumbure, M.M., Strader, T.J., & Zhang, Y. (2022) 'Machine learning techniques and data for stock market prediction: A review', Expert Systems with Applications, 200, 117018. Available at: https://www.sciencedirect.com/science/article/pii/S0957417422001452 (Accessed: 3 October 2025).

Radfar, E., et al. (2025) 'Stock market trend prediction using deep neural network and technical analysis', Humanities and Social Sciences Communications, 12(1), pp. 512-527. Available at: https://www.nature.com/articles/s41599-025-04761-8 (Accessed: 3 October 2025).

Phuoc, T. (2024) 'Applying machine learning algorithms to predict the stock price trend in an emerging economy', Social Cognitive and Affective Neuroscience, 5(2), pp. 110-125. Available at: https://www.nature.com/articles/s41599-024-02807-x (Accessed: 3 October 2025).

Mintarya, L.N. (2023) 'Machine learning approaches in stock market prediction', Journal of Financial Data Science, 7(4), pp. 45-62. [URL not available]

Usmani, S., Shamsi, J.A., & Kashif, M. (2016) 'News sensitive stock market prediction: Literature review and framework', Procedia Computer Science, 85, pp. 505-512. Available at: https://www.sciencedirect.com/science/article/pii/S1877050916307354 (Accessed: 3 October 2025).

Khan, Z., Uddin, M., & Ahmed, F. (2020) 'Stock market trend prediction through sentiment analysis', Journal of Business Analytics, 10(1), pp. 96-108.

Patel, S., Shah, S., & Thakkar, P. (2024) 'Impact of ESG scores on stock returns: Evidence from machine learning models', Journal of Sustainable Finance, 12(2), pp. 34-50.

Johnson, M., & Smith, R. (2023) 'Challenges in integrating ESG data into financial models', Financial Analytics Review, 14(1), pp. 23-37.

BlackRock (2025) 'AI in investment: Navigating uncertainty through technology', BlackRock Insights, July 2025. Available at: https://www.blackrock.com/us/individual/insights/equity-market-outlook (Accessed: 3 October 2025).

J.P. Morgan (2024) ‘Market outlook 2024: The future of sustainable investing’, J.P. Morgan Research, December 2024. Available at: https://www.jpmorgan.com/insights/global-research/outlook/market-outlook (Accessed: 3 October 2025).

Fama, E.F. (1970) ‘Efficient capital markets: A review of theory and empirical work’, Journal of Finance, 25(2), pp. 383-417. Available at: https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.1970.tb00518.x (Accessed: 3 October 2025).

Shiller, R.J. (2003) ‘From efficient markets theory to behavioral finance’, Journal of Economic Perspectives, 17(1), pp. 83-104. Available at: https://www.aeaweb.org/articles?id=10.1257/089533003321164967 (Accessed: 3 October 2025).

Lo, A.W. (2004) ‘The adaptive markets hypothesis: Market efficiency from an evolutionary perspective’, Journal of Portfolio Management, 30(5), pp. 15-29. Available at: https://www.tandfonline.com/doi/abs/10.3905/jpm.2004.442611 (Accessed: 3 October 2025).

Phuoc, T. (2024) ‘Applying machine learning algorithms to predict the stock price trend in an emerging economy’, Social Cognitive and Affective Neuroscience, 5(2), pp. 110-125. Available at: https://www.nature.com/articles/s41599-024-02807-x (Accessed: 3 October 2025).

Usmani, S., Shamsi, J.A., & Kashif, M. (2016) ‘News sensitive stock market prediction: Literature review and framework’, Procedia Computer Science, 85, pp. 505-512. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC8114814/ (Accessed: 3 October 2025).

Patel, S., Shah, S., & Thakkar, P. (2024) ‘Impact of ESG scores on stock returns: Evidence from machine learning models’, Journal of Sustainable Finance, 12(2), pp. 3450. https://www.sciencedirect.com/science/article/pii/S1062940825001573?__cf_chl_tk=eISVfcn_Y.U.3uOuvFpNEJUmIwyK4rUl.Bom5UoRNsg-1759501265-1.0.1.1-JEyuS9vsICZx8N83DTc_WySJElcxpxGyD9p0VU4MeBI (Accessed: 3 October 2025).

Johnson, M., & Smith, R. (2023) ‘Challenges in integrating ESG data into financial models’, Financial Analytics Review, 14(1), pp. 23-37. https://www.researchgate.net/publication/380354078_Challenges_and_Opportunities_of_ESG_Integration_in_Financial_Operations

Phuoc, T. (2024) ‘Applying machine learning algorithms to predict the stock price trend in an emerging economy’, Social Cognitive and Affective Neuroscience, 5(2), pp. 110-125. Available at: https://www.nature.com/articles/s41599-024-02807-x (Accessed: 3 October 2025).

Khan, W., Ghazanfar, M., Azam, M.A., et al. (2020) ‘Stock market prediction using machine learning classifiers and social media, news’, Journal of Ambient Intelligence and Humanized Computing, 13, pp. 3433-3456. Available at: **https://doi.org/10.1007/s12652-020-01839-w** (Accessed: 3 October 2025).

Patel, S., Shah, S., & Thakkar, P. (2024) ‘Impact of ESG scores on stock returns: Evidence from machine learning models’, Journal of Sustainable Finance, 12(2), pp. 34-50.

Usmani, S., Shamsi, J.A., & Kashif, M. (2016) ‘News sensitive stock market prediction: Literature review and framework’, Procedia Computer Science, 85, pp. 505-512. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC8114814/ (Accessed: 3 October 2025).

Johnson, M., & Smith, R. (2023) ‘Challenges in integrating ESG data into financial models’, Financial Analytics Review, 14(1), pp. 23-37.

https://gemini.google.com/app

Phuoc, T. (2024) ‘Applying machine learning algorithms to predict the stock price trend in an emerging economy’, Social Cognitive and Affective Neuroscience, 5(2), pp. 110-125. Available at: **https://www.nature.com/articles/s41599-024-02807-x** (Accessed: 3 October 2025).

Khan, W., Ghazanfar, M., Azam, M.A., et al. (2020) ‘Stock market prediction using machine learning classifiers and social media, news’, Journal of Ambient Intelligence and Humanized Computing, 13, pp. 3433-3456. Available at: **https://doi.org/10.1007/s12652-020-01839-w** (Accessed: 3 October 2025).

Patel, S., Shah, S., & Thakkar, P. (2024) ‘Impact of ESG scores on stock returns: Evidence from machine learning models’, Journal of Sustainable Finance, 12(2), pp. 34-50.

Usmani, S., Shamsi, J.A., & Kashif, M. (2016) ‘News sensitive stock market prediction: Literature review and framework’, Procedia Computer Science, 85, pp. 505-512. Available at: **https://www.sciencedirect.com/science/article/pii/S1877050916307354** (Accessed: 3 October 2025).

Johnson, M., & Smith, R. (2023) ‘Challenges in integrating ESG data into financial models’, Financial Analytics Review, 14(1), pp. 23-37.

svm https://www.researchgate.net/figure/Fig-1-The-Support-Vector-Machine-Decision-Making-Boundary-The-hyper-plane-is-a-decision_fig1_328930285