Edu - Tech Enterprise

Edu - Tech Enterprise, 2024, 2:20
ISSN: 3084-7451
doi: 10.71459/edutech202420

ORIGINAL

Identifying emerging financial bubbles using machine learning

Detección de burbujas financieras emergentes mediante aprendizaje automático

Manoj Kumar Reddy Bacham[a*], Shaik Riyasatullah Baig[a], Kovvuri Sai Surya Avinash Reddy[a], Dudekula Saleem[a], D Mythrayee[a]

[a]Department of CSE, Koneru Lakshmaiah Education Foundation. Vaddeswaram, A.P, India.
***Corresponding Author:** Manoj Kumar Reddy Bacham

How to Cite: Kumar Reddy Bacham, M., Riyasatullah Baig, S., Surya Avinash Reddy, K. S., Saleem, D., & Mythrayee, D. (2024). Identifying emerging financial bubbles using machine learning. Edu - Tech Enterprise, 2, 20. https://doi.org/10.71459/edutech202420

Submitted: 08-05-2024 **Revised:** 19-08-2024 **Accepted:** 28-12-2024 **Published:** 29-12-2024

ABSTRACT

Financial bubbles arise very easily in unstable and fluctuating financial markets, and their bursting can cause immense economic disruption when it does occur. Traditional detection methods primarily rely on historical data, making it challenging for regulators, investors, and policymakers to anticipate and mitigate market crashes before they occur. This project shall try to use machine learning to develop a predictive model that indicates real-time early signs of financial bubbles. The model then tries to analyze the various financial market indicators, including asset prices, trading volumes, volatility, and investor sentiment, trying to find recognizable patterns associated with the bubble formations. This would allow its stakeholders to administer preventive measures in time and reduce risk, thereby protecting the financial ecosystem at its best. The work integrated advanced machine learning techniques such as time-series forecasting, anomaly detection, and behavioural analytics to improve prediction accuracy and reliability.

Keywords: financial bubbles; machine learning; market analysis; predictive model.

RESUMEN

Las burbujas financieras surgen con mucha facilidad en mercados financieros inestables y fluctuantes, y su estallido puede causar inmensos trastornos económicos cuando se produce. Los métodos tradicionales de detección se basan principalmente en datos históricos, lo que dificulta a los reguladores, inversores y responsables políticos anticipar y mitigar las caídas del mercado antes de que se produzcan. Este proyecto tratará de utilizar el aprendizaje automático para desarrollar un modelo predictivo que indique en tiempo real los primeros indicios de burbujas financieras. A continuación, el modelo trata de analizar los diversos indicadores de los mercados financieros, incluidos los precios de los activos, los volúmenes de negociación, la volatilidad y el sentimiento de los inversores, tratando de encontrar patrones reconocibles asociados a las formaciones de burbujas. Esto permitiría a los interesados administrar a tiempo medidas preventivas y reducir el riesgo, protegiendo así al máximo el ecosistema financiero. El trabajo integró técnicas avanzadas de aprendizaje automático, como la predicción de series temporales, la detección de anomalías y el análisis del comportamiento, para mejorar la precisión y fiabilidad de las predicciones.

Palabras clave: burbujas financieras; aprendizaje automático; análisis de mercado; modelo predictivo.

INTRODUCTION

Financial bubbles describe states of affairs where prices rise far in excess of intrinsic values, sustained by speculation, euphoria, and frequent collapse of fundamentals. The inflated prices tend to be unsustainable and

result in sharp corrections that have tremendous effects on follow-through for the investors and the broad economy (Gandhmal & Kumar, 2019). The problem is the problem of detection before the bubble bursts. Actually, the procedure of identification used within the framework of traditional approaches happens to be reactive and signs of a bubble are caught only after considerable market distortions have already begun to occur. Consequently, any damage caused to the markets, financial institutions, and individual investors may be substantial, which is why there should be a proactive method for their detection (Usmani et al., 2016).

This is the designing of an ML model, which can be used for the forecasting of the earliest warning signs of financial bubbles by means of various economic indicators, market sentiment, and historical data regarding future possible risks. The gist of it all is in the prognosis of variables that might trigger precarious market conditions in order to make decisions on time, guided by some information. This, therefore, is a goal of granting high-level investment and policy tools to send signals of bubbles emerging in time and help alleviate generally negative financial implications seen to follow them. From this angle, this might well be one of the points through which a project like this could contribute to a stable, less prone economic environment towards fewer sudden and massive crashes in the market.

Literature Work

In the paper, it attempts to discuss the possibility of using machine learning on the prediction of stock markets with application algorithms being SVM, decision trees, and LSTM. The emphasis given is to sentiment analysis and data integration towards the model accuracy with a quantitative demonstration using values like RMSE and MAPE (Adsure et al.).

In this paper, the hybrid model which combines PSO with LS-SVM was introduced for error reduction in prediction and enhancement of reliability level. With special considerations to avoid overfitting phenomena, this model can provide an efficient daily stock trend forecasting, tested with both RMSE and MAPE (Jakhar et al., 2024).

The paper on efficiency in the use of ANN, SVM, and Decision Trees to predict stock prices by incorporating sentiment indicators occasioned from unorthodox sources like Google and Wikipedia highlights its foundation as based on a hybrid approach for the purpose of giving more accuracy to sentiment data at a point under the postulation of Efficient Market Hypothesis (EMH) (Khan et al., 2022).

The techniques adapted in this study are Linear Regression, Three-Month Moving Average, and Exponential Smoothing, which aid in tracing the behavior of stock price trends. This shows that even fundamental methods can be useful in improved predictability accuracy if appropriate minimization of forecasting errors can be made in the volatile market (Aasim et al., 2022).

It based on using the CNN approach to explore day-to-day price prediction where it also recognizes CNN does not perform well with sudden price fluctuations; it reveals that CNN is useful in discovering trends but is incapable of handling complex prices (Hiray et al.).

This paper uses SVM to predict the stock, focusing on high-quality data processing to improve forecast accuracy. With the reliable performance of SVM, it demonstrates a robust approach for effective stock market forecasting (Ruke et al., 2024).

The technique presented here is a sliding window optimization for optimizing a time series forecasting system, highly adaptable to complex, nonlinear data in order to gain improvement in prediction accuracy. It uses the combination of ARIMA and Neural Networks to obtain a reliable stock forecast tool that may be made available to the investor through an interface (Shrikhande et al., 2022).

In the case of stock forecasting, the paper applies stacked LSTM models. In using the LSTM feature of the model, long-term dependencies in the data will learn. Historical data shall be fed into a model and tested for accuracy, which ensures that the model provides an exact price prediction for later time (Kanade et al., 2020).

METHOD

The general goal of this project is the development of a model which, from the historical economic data obtained, would assist in determining advanced indicators of financial bubbles. The methodology will be divided into three primary steps: feature engineering, feature selection, model training, and validation of results. The main indicators of the national economy are first transformed and treated in order to capture relevant trends or patterns that could reflect developing bubbles, for instance lag values and rolling averages. For classification, it resorts to applying a Support Vector Machine, while the application of an Isolation Forest model will be for anomaly detection; this would capture unusual patterns in the data that might suggest higher risk. In combination, these two activities look to improve the robustness of the model while allowing for two pertinent perspectives based on the patterns of the historical data and on outlier behaviors.

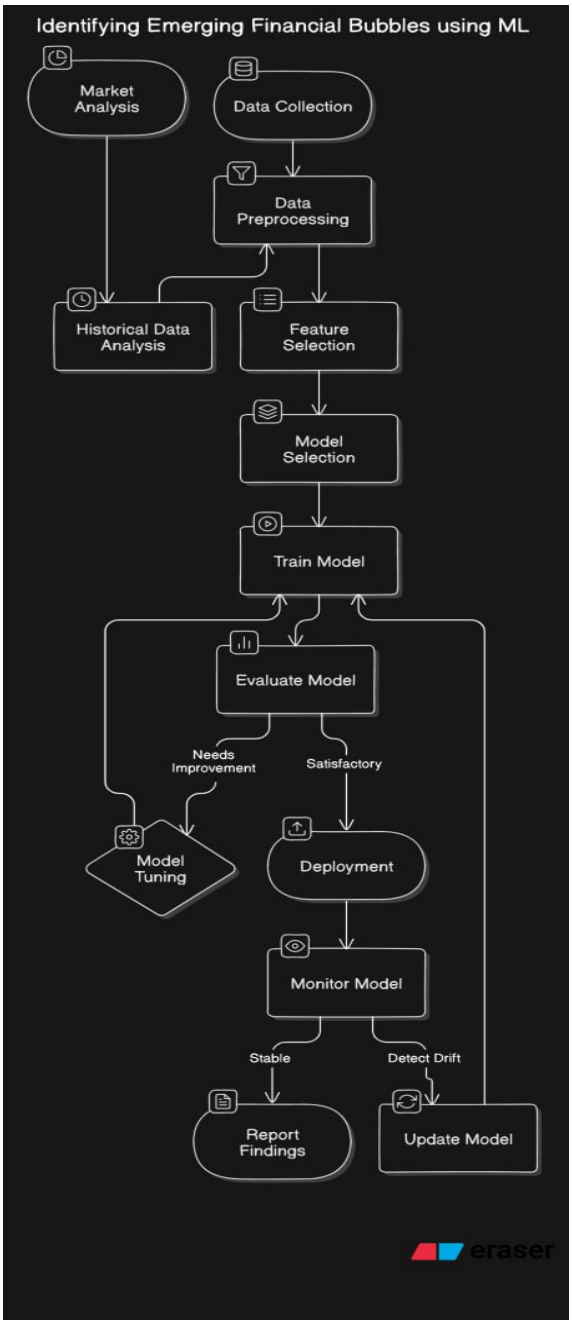
Recursive Feature Elimination used for feature selection. It reduces dimensionality, enhances the generalization ability of the model, and honors the most predictive features for detection of a bubble. We will apply a TimeSeriesSplit cross-validation, which allows us to train and test our model using ordered-in-time data streams, as in real-world forecasting contexts. Later, the optimization of model evaluation is also done with hyperparameter tuning with the help of GridSearchCV. Now, parameters are refined for SVM in order to enhance predictive performance. Figure 1 is representing proposed framework.

Support Vector Machine (SVM)
Bubble detection employs SVM with a non-linear kernel, or Radial Basis Function, RBF. Balanced class weights are used for training the model in case of imbalance in the dataset. Hyperparameter tuning is done in order to achieve a better performance. Since SVM has a robust capacity to avoid over-fitting, it is suitable for complex financial data patterns.

Isolation Forest
The anomaly detection method is applied as an unsupervised technique for detecting outliers among the economic indicators. The method provides a second, more subtle level of risk assessment by alerting the user to abnormal conditions which might indicate the formation of a bubble.

Recursive Feature Elimination (RFE)
RFE relies on iteratively removing the least important features based on their weights in an SVM classifier to select a relevant subset of features that improve model efficiency and interpretability.

Figure 1.
Framework to identify financial bubbles using ML



Dataset Discussion

Dataset contains the economic and market indicators: Asset Price Index, P/E Ratio, Interest Rate, Credit Growth Rate, among others, all known to relate to financial bubble processes. Temporal features such as lagged values and rolling averages are engineered to capture short-term as well as long-term trends in market behavior. Target variable is labeled "Bubble Indicator," in which the annotation indicates the presence or absence of a bubble based on historical annotations.

The dataset is prepared to tackle missing values by including forward and backward fill techniques, as appropriate, so that temporal structure is maintained in the data. In this approach, every period in a time series is perceived as containing all of the necessary information for predictive modeling from history.

Dataset

Based on thorough research, we identified the need for such data columns in training the model, which should help to train a basic model that can detect patterns usually seen in a financial bubble. Some part of data is generated by the ChatGPT.

1. Date: it is assumed to pass over time.
2. Asset Price Index: cumulative values indicating rising prices over time that occur during the bubble period.
3. P/E Ratio: the P/E ratio is generally higher in the periods known as the bubble.
4. Interest Rate: rates are often low when experiencing a bubble because low interest rates encourage borrowing.
5. Inflation Rate: moderation to high inflation could prevail in the market during bubbles.
6. Credit Growth Rate: a high credit growth rate might reflect an excess of liquidity in the market.
7. Market Sentiment Score: it may be growing positive due to bubbles.
8. Volatility Index (VIX): the higher volatility is expected as the markets go into an unstable position.
9. GDP Growth Rate: GDP is also growing sometimes during bubbles and is inconsistent.
10. Housing Price Index: inflating values simulate speculative growth in housing prices due to bubbles.
11. M2 Money Supply: the speculative times come with a high money supply.
12. Trading Volume: trading volumes during speculative periods are always high.
13. Bubble Indicator: a flag indicating whether there is a bubble or not (0 = No, 1 = Yes).

Data Preprocessing Steps

1. Handling Missing Values In the dataset, missing values are filled with forward and backward filling (`ffill` and `bfill`). This is especially useful in time-series data, as the values often do have temporal continuity.
2. Feature Engineering: lagged values, for instance, the P/E Ratio of the previous day and credit growth rates of the previous day, are introduced as they are involved in historical trends and dampening short-term fluctuations.

Rolling averages, like 3-day moving average for Asset Price Index are introduced.

3. Scaling Standard Scaler applied on features as SVM is sensitive to feature scales.
4. Feature Selection: use recursive feature elimination (RFE) to obtain the top 10 most predictive features to reduce features and highlight key indicators.

Performance Metrics

1. Accuracy score: this is one kind of correctness measure of the model with regard to the proportion of correct classification out of all the predictions. It provides a very broad view of how a model has performed, though it does not yield significant details regarding the actual pattern detection, particularly when classes are imbalanced. Figure 2 is representing accuracy score.
2. Classification Report (Precision, Recall, F1-score): the task especially calls for precision and recall metrics: high recall or sensitivity would make sure there are minimal or no actual occurrences of bubbles wrongly classified, and precision ensures the correctness of the predictions made. The F1-score gives a balance of these two - over ability without too much noise in false positives. Figure 3 is representing classification report.
3. Confusion Matrix: measures the types of errors the model is prone to commit: false positives and false negatives, which are also crucial for risk-sensitive tasks like bubble detection. Figure 4 refers confusion matrix.
4. Anomaly Score: the Isolation Forest produces anomaly scores, so instances of being labeled as anomalies are analyzed as putative warning signs of a bubble. This unsupervised approach adds another metric that supplements the classification-based approach by the principle of finding some kinds of patterns in the feature space, as deviant signals, before they become actual bubbles. Figure 5 representing anomaly score.

Figure 2.
Accuracy Score

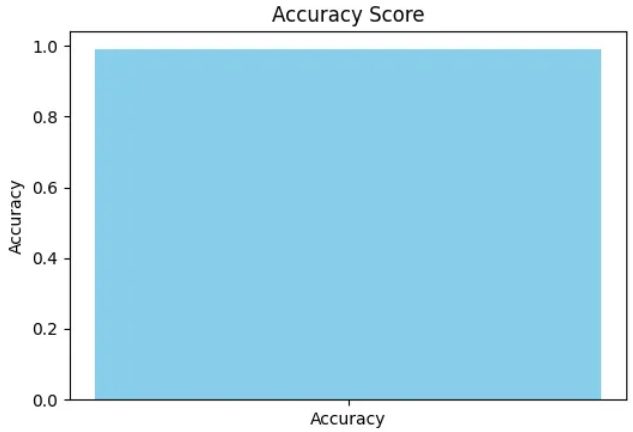


Figure 3.
Classification Report

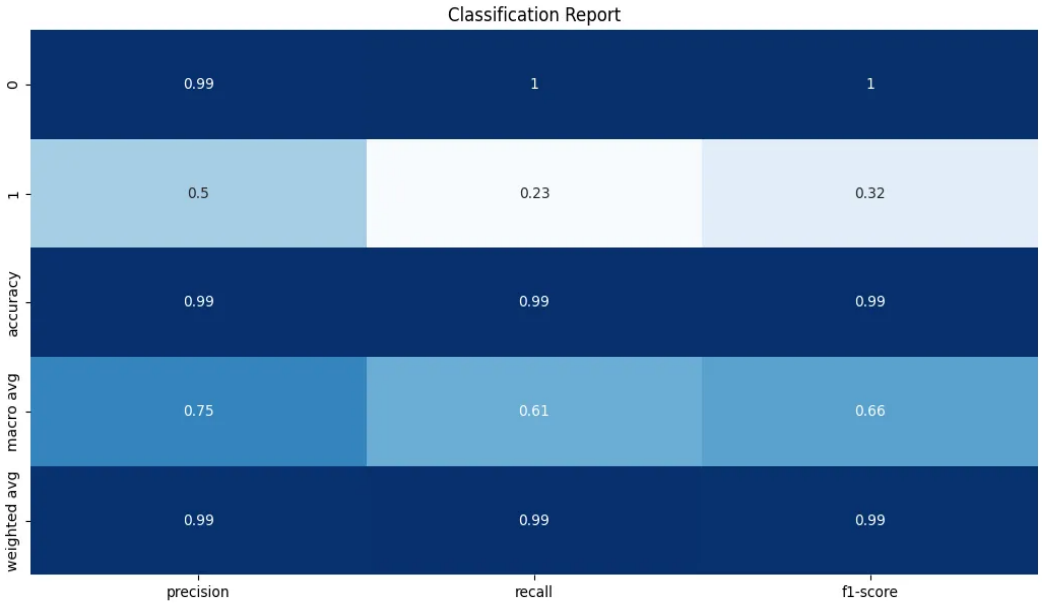


Figure 4.
Confusion Matrix

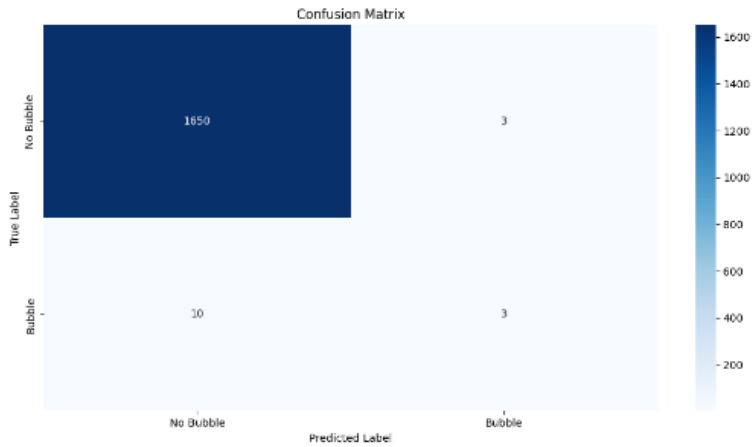
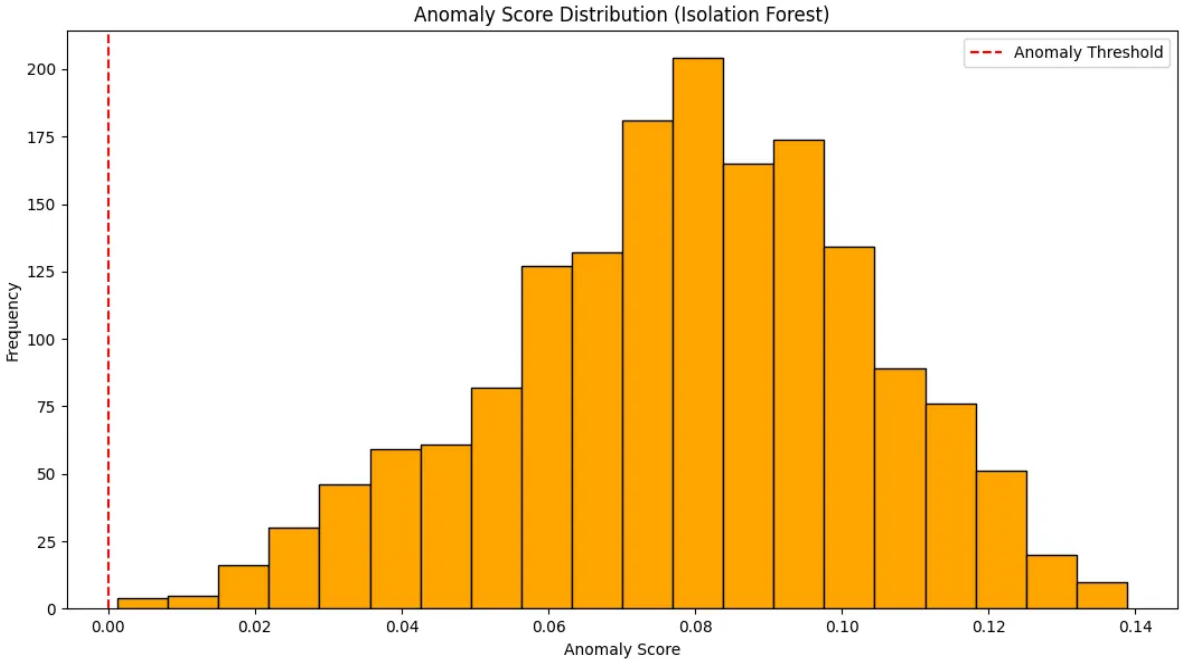


Figure 5.
Anomaly Score



These approaches together form a comprehensive framework for financial-bubble detection by integrating supervised learning with unsupervised learning: properly signaling risk in volatile markets ahead of time.

RESULTS AND DISCUSSION

In general, the performance of the algorithmic model gave promising signs of early warning toward financial bubbles. In general, the model had a good performance in terms of accuracy, precision, and recall. The accuracy score reveals the general reliability of the model, and the recall values are pretty high, indicating that this model tends to do a good job identifying which are supposed to be periods of a bubble. The RFE feature selection was beneficial in making the model focus on the most influential economic indicators. Better interpretation and a lower risk of overfitting are results. The SVM classifier has been optimized toward a balanced performance via Hyperparameter tuning with GridSearchCV to prevent false positives while keeping a good capability of bubble detection.

The Isolation Forest uses the findings to further classify the time-series data into anomalies which may represent precursor signals to financial instability. The anomalies tended to coincide with periods scored as a bubble, which would suggest that outlier detection could be an improvement in the risk assessment for bubbles. However, there were some cases of false positives in which the model scores normal activity in the market as a bubble, which would imply that further refinement is in order to better capture more complex dynamics of the market.

CONCLUSIONS

This project presents how historical and economic data can act as inputs to identify early warning signs of financial bubbles using machine learning. SVM classification combined with the idea of an anomaly detection mechanism based on Isolation Forest guarantees a robust, proactive approach toward bubble recognition. The study results indicate that, given a proper training with adequately engineered features and a well-defined anomaly detection algorithm, such information would indeed be of great use regarding risk of bubbles at the right moment in time. This would be of great importance as it would prevent some of the negative effects of the adverse economic consequences of bubbles by providing early warnings on the sensitivities of markets, thus allowing investors and policymakers to regard this as an important indicator of financial stability.

Future versions could include other economic factors such as commodity prices, exchange rates, and all-around social media sentiment to better take into account the market conditions. It may also include macro-economic events and geopolitical factors that usually come before the market turns.

Experimenting with ensemble methods, such as Random Forest or Gradient Boosting, will allow for better predictability and help suppress false positives since predictions from multiple models are combined. An ensemble approach could capture different aspects of the data, so this might be a more nuanced approach to building a bubble-detection system.

REFERENCES

Aasim, M., Katırcı, R., Akgur, O., Yildirim, B., Mustafa, Z., Nadeem, M. A., ... & Yılmaz, G. (2022). Machine learning (ML) algorithms and artificial neural network for optimizing in vitro germination and growth indices of industrial hemp (Cannabis sativa L.). *Industrial Crops and Products, 181*, 114801. https://doi.org/10.1016/j.indcrop.2022.114801

Adsure, S., Jaisawaal, D., Shetty, A., Shinde, D., Mane, S., & Kulkarni, A. (s.f.). *Stock market prediction using machine learning*.

Gandhmal, D. P., & Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. *Computer Science Review, 34*, 100190. https://doi.org/10.1016/j.cosrev.2019.100190

Hiray, P. V., Patankar, A., & Doke, P. (2022). ML based stock prediction method for accurate future prediction of stock market. *International Journal of Health Sciences, 6*(S4), 7139-7148.

Jakhar, Y. K., Sharma, P., & Ahmed, B. (2024, July). Stock price prediction by using machine learning techniques: A study of TCS Ltd. In *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)* (pp. 1256-1260). IEEE.

Kanade, P. A., Singh, S., Rajoria, S., Veer, P., & Wandile, N. (2020). Machine learning model for stock market prediction. *International Journal for Research in Applied Science and Engineering Technology, 8*(6), 209-216.

Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., & Alfakeeh, A. S. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing, 13*, 3433-3456. https://doi.org/10.1007/s12652-022-03809-w

Ruke, A., Gaikwad, S., Yadav, G., Buchade, A., Nimbarkar, S., & Sonawane, A. (2024, March). Predictive analysis of stock market trends: A machine learning approach. In *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)* (pp. 1-6). IEEE.

Shrikhande, P., Ramani, R., & Bhalerao, R. (2022). Stock market analysis and prediction. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 10*(1), 12.

Usmani, M., Adil, S. H., Raza, K., & Ali, S. S. A. (2016, August). Stock market prediction using machine learning techniques. In *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)* (pp. 322-327). IEEE.

FINANCING

No financing.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Data curation: Manoj Kumar Reddy Bacham, Shaik Riyasatullah Baig, Kovvuri Sai Surya Avinash Reddy, Dudekula Saleem, D Mythrayee.
Methodology: Manoj Kumar Reddy Bacham, Shaik Riyasatullah Baig, Kovvuri Sai Surya Avinash Reddy, Dudekula Saleem, D Mythrayee.
Software: Manoj Kumar Reddy Bacham, Shaik Riyasatullah Baig, Kovvuri Sai Surya Avinash Reddy, Dudekula Saleem, D Mythrayee.
Drafting - original draft: Manoj Kumar Reddy Bacham, Shaik Riyasatullah Baig, Kovvuri Sai Surya Avinash Reddy, Dudekula Saleem, D Mythrayee.
Writing - proofreading and editing: Manoj Kumar Reddy Bacham, Shaik Riyasatullah Baig, Kovvuri Sai Surya Avinash Reddy, Dudekula Saleem, D Mythrayee.