Unmasking Financial Deception: Leveraging Machine Learning for Risk Analysis and Fraud Detection in Financial Markets

Yashika Garg
CSE, IGDTUW
New Delhi, India

Karuna Kadian
CSE, IGDTUW
New Delhi, India

Varda Mahajan*
CSE, IGDTUW
New Delhi, India

Somya
CSE, IGDTUW
New Delhi, India

***Abstract*—Throughout the past two centuries, the stock market has served as a vital source of income for countless individuals, and its significance in generating wealth is expected to endure in the years ahead. Many financial analysts have long aspired to predict the ebbs and flows of the stock market, but their attempts have yielded limited success. Thanks to the remarkable advancements in technology, characterized by exponential increases in processing power, enhanced storage capabilities, and sophisticated algorithms, the prospect of achieving this goal is now more promising than ever. However, breaking the stereotype that the stock market is exclusive to those well-versed in finance necessitates the development of a novel solution. This solution comprises of the application of different machine learning algorithms to inform interested users about the level of associated risks and frauds in the market. It uses the ESG risk dataset for training the model, risk analysis and fraud detection to overcome the existing barrier and ensure that the benefits are accessible to a wider audience.**

***Index Terms*—Ensemble learning method, Financial risk assessment, Fraud detection, Decision trees, Predictive accuracy, Overfitting reduction, Anomaly detection, Interpretability sacrifice**

I. Introduction

The stock market, as one of the pillars of global financial system, provides opportunities for investment and wealth creation. The primary function of the stock market is to facilitate the allocation of capital [9]. The central idea to successful and fortunate stock market forecast is to achieve the best results after training minimum data as input and the least complex stock market model [4].

Companies looking to raise funds for various purposes, such as research, expansion and development, or debt repayment, can issue shares of stock to the public [20]. However, it also presents significant challenges in terms of risk and fraud detection. In recent years, the application of Machine Learning (ML) and Artificial Intelligence (AI) techniques has revolutionized the way we approach risk management and fraud detection in the stock market. These advanced technologies have empowered financial institutions and regulatory bodies to enhance their surveillance capabilities and respond more effectively to emerging threats [12].

Although extensive research has been done in fraud analysis and risk prediction using ML, there have been no significant researches that include both risk analysis and fraud analysis using the same ML model. The S&P 500 index data set was used which was ultimately successful in performing the desired tasks. Moreover, there have not been much research that was able to include multiple algorithms for finding the accuracy of the model [16]. The numerous algorithms that were used for the research were able to give very accurate and precise results with a high degree of accuracy. This made our research distinctive and impactful.

The research paper is well-structured and follows a logical progression from introduction to conclusion. The following section consists of the literature review II which is comprehensive, covering relevant studies and encompassing research done in the concerned field. The methodology III section is detailed and transparent, allowing for replication of the study. The methods that were used to analyze risk and predict fraud are logistic regression (III-A), KNN (III-B), naive bayes(III-C), decision tree (III-D), random forest(III-E) and boosting(III-F). Implementation IV consists of the dataset used and describes the process of applying ML models that leads to finding the accuracies of the multiple models used for risk analysis and fraud detection. Results V are presented clearly with the appropriate use of tables and figures, and the discussion thoughtfully interprets these findings in the context of existing literature. The conclusion VI succinctly summarizes the research, and its implications, and suggests directions for future studies. References are complete and properly formatted. Overall, the paper flows logically and maintains coherence throughout, making it easy to follow and understand.

II. Literature Review

Extensive research has been conducted within the domain of stock market analysis, yet achieving the desired levels of precision has remained elusive. ML techniques have been employed not solely based on past stock prices but have also been utilized with additional factors like trading volume and average variance to enhance predictive accuracy [9]. ML has seen a significant evolution in the areas of risk analysis and fraud detection. Early techniques utilized statistical models and rule-based systems. With the introduction of ML, methods like SVM, Random Forest, and deep learning, detection capabilities were greatly improved [17]. To improve fraud detection systems, recent trends emphasize interpretability and transfer learning. ML is a subset of AI that leverages algorithms and statistical models to enable computers to learn from and make predictions or judgements based on data. In the context of the stock market, ML has emerged as a powerful tool for identifying and mitigating risks while detecting fraudulent activities [13]. Risk analysis is a vital practice in the financial markets, essential for assessing and managing various types of risks that can affect investment portfolios, market stability, and overall financial well-being. With the advent of ML models and advanced data analytics techniques, risk analysis has entered a new era of sophistication and accuracy. Fraud detection is a critical aspect of maintaining the integrity and security of financial markets. It involves identifying and preventing fraudulent activities such as insider trading, market manipulation, Ponzi schemes, and unauthorized access to sensitive information. ML offers a powerful set of tools and techniques for fraud detection in financial markets which includes Model Training, data processing, Anomaly detection etc. ML has transformed the financial industry by enabling the analysis of large amounts of data, enhancing prediction accuracy, and automating complex tasks [18]. By leveraging ML techniques, financial professionals can make more informed decisions, manage risks effectively, and adapt to the ever-changing dynamics of financial markets [7].

III. METHODOLOGY

The dataset exclusively showcases companies from the S&P 500 index.The dataset has been utilized to gain insights into the ESG performance and risk profiles of some major corporations [15]. The training data set would include risk and controversy/fraud levels in these major corporations. The following methods have been used to analyze risk and predict frauds for different corporations-

A. Logistic Regression

Logistic Regression is a factual modeling procedure which is significant in money related markets for risk assessment and fraud detection. It measures the likelihood of double results based on autonomous factors [19]. In risk ananlysis, it analyzes chronicled information to assess the probability of occasions like showcase downturns or defaults. In fraud detection, it evaluates exchange information to recognize suspicious exercises. Calculated relapse offers interpretability and proficiency, making it a profitable apparatus. In any case, it expect linearity and may be sensitive to exceptions. It uses the logistic function, called as sigmoid function, to convert a linear combination of independent variables into values between 0 and 1. Here's the general formula:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}} \quad (1)$$

Where: The probability that the binary outcome Y will be 1 (success) given the values of the independent variables X is expressed as P(Y=1—X).

The natural logarithm's base is e, or roughly 2.71828.

- β_0 known as intercept or consistent term.
- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients related to the independent variables $X_1, X_2, \ldots, X_n$, respectively.

In simple terms, the logistic regression formula calculates the probability that the dependent variable (Y) equals 1 based on a linear combination of the independent variables (X) transformed by the logistic function

B. KNN

k-Nearest Neighbors (KNN) is a supervised ML algorithm that can be utilised for risk assessment and fraud detection in financial market. The algorithm is a non-parametric(not make any assumption on underlying data), supervised learning classifier, which predict new data, by finding similarity with existing data points in training dataset [1].

KNN identifies patterns that may be associated with risk or fraud by analyzing historical data. When the algo arrives across any new change , similarities to the old are important identifies previously collected and presented data points corresponding risk levels, making it easier to determine real-time potential risks or fraudulent activities [14]. KNN can be effectively used in the regularly changing financial market environment, where patterns may evolve over time. For optimal application, careful consideration of parameter tuning and feature selection is necessary.

Formula for the K-Nearest Neighbors (KNN) set of rules It may be explained as follows. Another statistics point x, .

1) Calculate the distance among x and all different information points In the dataset.
2) Identify the K-adjacent facts points primarily based at the calculation.
3) Give the new data point, the class in which majority of k nearest neighbours are present .

In mathematics if we consider a data point as (x1, x2, . . . , xn) and d(x, y) are distance metrics, ma- . mula To use KNN to predict the learning of the new data point It can be expressed as follows:

$$\text{Prediction for } \mathbf{x} = \text{mode}\left(\{y_i \mid i = 1, 2, \ldots, K\}\right) \quad (2)$$

x = prediction for the property (this — i = 1, 2, . . . , K) . where yi is the category label of the i-th closely neighbor.

C. Naïve Bayes

Naive Bayes works on the hints of Bayes' theorem, which calculates the probability of an event happening based on in advance know-how. In this context, naive Bayes assesses the probability of a financial transaction being volatile or fraudulent given positive functions. It assumes a point distribution of freedom, which simplifies calculations and yet generally does not hold for real-world conditions. Naive Bayes is specifically compelling in scenarios with a big quantity of highlights and can deliver speedy and unique expectations. The equation for the naive Bayes calculation can be depicted as takes after: Given a highlights vector x = (x1, x2, . . . , xn) Naive Bayes is particularly compelling in situations with a big range of highlights and may supply rapid and particular expectations. It may require cautious highlight choice and pre handling for ideal execution. In spite of its 'naive' presumption, Naive Bayes remains a profitable instrument in the weapons store of monetary chance administration and extortion anticipation methodologies. The equation for the Naive Bayes calculation can be depicted as takes after: Given a highlights vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ representing the attributes of a data point, and a class variable C,

1. Calculate the conditional probability of the class variable C, given the feature vector x, denoted as P(C $| x$).

2. Using Bayes' theorem, this can be expressed as:

$$P(C \mid x) = (P(x \mid C) * P(C))/P(x) \tag{3}$$

, where:

- $P(x \mid C)$ is the chance of encountering x given class C.
- P(C) is the earlier probability of class C, - P(x) is the probability of observing the feature vector x.

The "naive" assumption in Naive Bayes simplifies the computation by assuming that the highlights are conditionally independent given the class variable C:

$$P(\mathbf{x} \mid C) = P(x_1 \mid C) \cdot P(x_2 \mid C) \cdot \ldots \cdot P(x_n \mid C) \tag{4}$$

D. Decision Tree

Decision trees use feature-based data splitting to enable them to make well-informed decisions. At danger evaluation, decision trees examine past financial data while taking into account economic indicators, business performance, and market trends [21]. This makes it conceivable to distinguish conceivable dangers like defaults or downturns within the advertise. They survey characteristics such as the volume, recurrence, and design of exchanges. They are delicate to both nonstop and categorical information, which makes them valuable in financial matters. The decision tree algorithm is based on a series of conditional statements that help in making decisions. Here is a description of the same:

Given a dataset with features $(X_1, X_2, \ldots, X_n)$ and a target variable Y, a decision tree repetitive partitions the information based on highlights values.The edges of the tree indicate potential outcomes, and each hub in the tree symbolises a decision point. The best feature (X_i) and matching threshold (X_i) that maximises information gain or Gini impurity reduction are chosen by the algorithm at each node. The definition of the decision rule is:

In the event where (X_i) is less than (X_i), then Proceed to the left child hub. Proceed to the appropriate child node. The aforementioned procedure carries on until a halting requirement is satisfied, such as a least number of samples per leaf, a maximum depth, or when more splits do not significantly increase the reduction of impurities.Ultimately, the leaf nodes hold the projected values or class labels. Ultimately, the leaf nodes hold the projected values or class labels. The decision tree formula is a sequence of IF-ELSE statements based on feature values that produce a classification or prediction at the end.

1. Start by using the complete dataset as the root.

2. For each node, calculate the best split:

1) For each feature X_i, consider all possible thresholds θ.
2) Calculate the information gain or Gini impurity reduction for each threshold.
3) Select the feature and threshold that give greatest data pick up or Gini impurity decreases.

- Apply the decision rule:

$$\text{IF } X_i \leq \theta \text{ THEN Go to left child node}$$

$$\text{ELSE Go to right child node}$$

- Repeat steps 2 and 3 for each child node until a ceasing measure is met: label=
  a. Greatest profundity is reached.
  b. Minimum samples per leaf are met.
  c. No significant improvement in impurity reduction.
- Assign the predicted value or class label to each leaf node.

Decision trees are a fundamental part of ML and provide clear, interpretable models for risk assessment and fraud detection in the financial domain.

E. Random Forest

Using Bagging (Bootstrap Aggregating) at Random Forest is a potent commercial venture that capitalises on the knowledge of widely utilised techniques in financial risk assessment and fraud detection. To improve predictive accuracy and reduce overfitting, include many choice woods, each taught sequentially in subgroup information [10]. By a set of results in many many trees, it offers a terribly intense analysis, enabling intelligent choices [3]. Also, sweeping trees in detecting fraud the algorithm enables the capture of a wide range of transactions Characteristics. This allows accurate anomaly detection and certainly more likely to detect frauds. It is very effective in complex financial management reinforce its role as a key tool for reducing estimation and overfitting for fraud detection professionals with financial risk. In an informal way in bagging, the forest rules include the selection of more than one forest have been studied in different statistical subgroups Here's a description for the same:

1. For B trees in the ensemble:

a. Sample a bootstrap dataset (with replacement) from the original data.

b. Train a decision tree on the bootstrap dataset, examining arbitrary subset of characterstics of the data after each split.

c. Repeat steps (a) and (b) to construct B individual trees.

2. For a classification task, each tree "votes" for the class of the input data point. The class with the majority votes gets to be final prediction. For regression, the predictions are averaged.

3. The ensemble prediction is obtained by comibining the results from all individual trees.

Random Forest with Bagging combines the predictions of multiple trees, resulting in a powerful and precise model. It reduces overfitting and improves the overall performance compared to a single decision tree.

F. Boosting

Boosting is an ensemble learning strategy that helps improve fraud detection and financial risk assessment. It systematically merges several poor models, like decision stumps, with the goal of fixing the flaws in the previous model [11]. Boosting uses past financial data to find complex patterns related to business performance measurements, economic indicators, and market movements in risk assessment. Through the course of iterations, the model adjusts and improve its predictions thanks to this sequential learning approach. Increasing the weight given to incorrectly categorised data points in every iteration allows boosting to precisely identify possibly fraudulent activity. It is beneficial and works well in the dynamic and continuously changing financial markets because it has the ability to detect errors. Boosting technique trains multiple weak learners successively that is the result from one is used as input for next and it finally produces a string learner. This is an explanation of the same:

1. Create a dataset and assign equal weight to each data point.

2. For each iteration from $t = 1$ to T (where T is the total number of iterations), perform the following steps:

1) Train a weak learner using the current dataset.
2) Determine the weak learner's error rate on the dataset.
3) Calculate the weak learner's weight α in the final ensemble.
4) Adjust the weights of the data points that were misclassified, so they carry more weight in the next iteration.
5) Generate a new dataset with updated weights for the subsequent iteration.

3. Combine the predictions of the weak learners by summing their weighted outputs to obtain the final prediction.

Boosting concentrates on enhancing the model's performance by iteratively giving more weight to miscategorized data points, enabling the model to learn from its mistakes. The ultimate prediction is determined by the weighted sum of the weak learners' outputs.

IV. Implementation

The dataset S&P 500 ESG Risk Ratings exclusively showcases companies from the $S\&P$ 500 index [5] . It consists of various risk scores namely the governance, environment, social, controversy level and ESG risk scores. The dataset also includes other attributes like several employees, industry sector, etc. For the application of the ML algorithms, the dataset was processed.

The processing included the replacement of the NaN values with the mean value of the data points of the particular features, alphabetical data was replaced with ordinal numbers by Label Encoding and redundant columns were dropped. After the processing of the data, various ML algorithms have been applied to check the accuracy of the model. After completing the preprocessing we were finally able to get a data set that was ready for training and testing and we were successfully able to apply different algorithms for predicting the accuracy of our dataset. We used multiple algorithms such logistic regression, kNN, naive bayes, random forest ,boosting and decision trees. For the successful implementation of the model we choose relevant X and Y, where X was the dataset without 'ESG Risk Level' column for risk analysis and 'Controversy Level' for fraud detection whereas Y was the column that was dropped in the respective cases. The following step comprises of splitting the data into train and test in the ratio of 8:2 . We then did the model fitting and found the accuracy of the train and test data using different algorithms as mentioned above. This way we were able to find the algorithms that were most suitable for the dataset and gave the best results [8]. Having the final results of all the algorithms from best to worth, we were able to achieve astonishingly high levels of accuracy. Ultimately we were successful in training our data set to analyze risks and detect frauds for the taken S&P 500 index data set [6]. The whole process is described using a flowchart as shown in Fig. 1 .

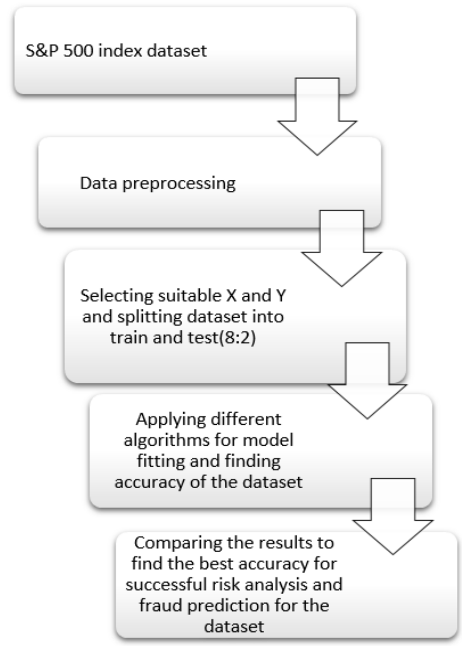

Fig. 1. Flowchart Showcasing Stepwise Implementation

The dataset lists ESG (Environmental, Social, and Governance) risk ratings for 503 companies listed in the S&P 500 index. Each row mentions the details of a company which is identified using its name. The dataset includes the following columns:

1) **Symbol**:Symbol of stocks for the company.
2) **Name**: the complete name of the company.

3) **Address**: Location of the headquater of the company.
4) **Sector**: The industrial segment under which the company falls.
5) **Industry**: The particular industry of the company.
6) **Full Time Employees**: Add up to no. of representatives working in the company.
7) **Description**: Some information regarding the activities of company.
8) **Total ESG Risk Score**: Score for the ESG risk in which higher scores indicate greater risks.
9) **Environment Risk Score**: Score telling environmental risks levels associated with the company.
10) **Governance Risk Score**: Score reflecting the governance risks.
11) **Social Risk Score**: Score reflecting the social risks.
12) **Controversy Level**: Qualitative categorisation of the controversy levels of a company (low,medium,high,...)
13) **Controversy Score**: Numerical score for controversy level.
14) **ESG Risk Percentile**: The percentile comparing the ESG risk score of a company to its peers.
15) **ESG Risk Level**: A categorical analysis of the composite ESG risk level (e.g., Low, Medium, High).

The dataset contains some missing values across various columns, with notable gaps in ESG risk scores and controversy levels for certain companies. The data provides a comprehensive overview of the ESG performance and associated risks for major companies, which can be useful for investors, analysts, and other stakeholders interested in sustainable and ethical business practices.

V. Result and Analysis

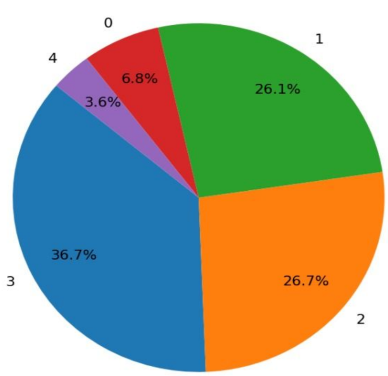

Fig. 2. Pie chart for ESG Risk Level

After testing all of the algorithms on the dataset, namely, S&P 500 ESG Risk Ratings, this study has been conducted to compare the accuracies attained in determining the fraud and the risk levels of different companies using various ML algorithms. The Fig. 2 shows the ESG risk levels for S&P 500 companies. The ESG risk level is negligible for 36.7% of the companies. Further, the ESG risk level is low for 26.1%, medium for 26.7%, high for 6.8% and severe for the rest 3.6% of the S&P 500 companies respectively. In a whole, there is ”negligible” risk to invest in more than one-third of the companies. The pie chart in Fig. 3 shows the controversy levels

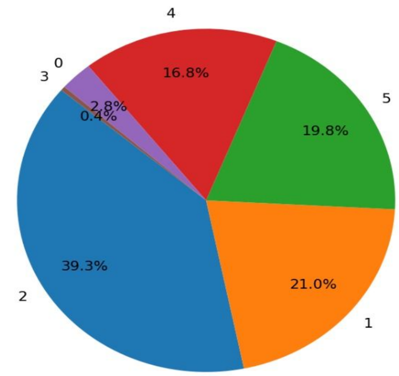

Fig. 3. Pie chart for Controversy Level

of S&P 500 companies. The controversy level is negligible for 2.8%, low for 21%, moderate for 39.3%, significant for 16.8%, high for 19.8% and severe for 0.4% of the S&P 500 companies respectively. Hence, the fraud rate is moderate for more than one-third of the companies. The Fig. 4 shows the

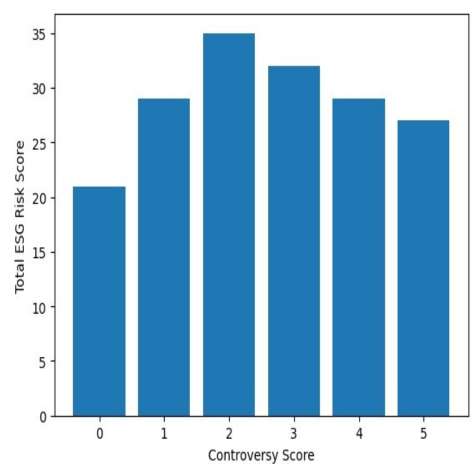

Fig. 4. Graph between controversy score and Total ESG Risk Score

relation between the ESG risk score and controversy score. ESG risk is maximum when controversy score is moderate

TABLE I
RISK ANALYSIS AND FRAUD DETECTION ACCURACIES

	RISK ANALYSIS ACCURACIES		FRAUD DETECTION ACCURACIES	
Algorithms	TRAIN	TEST	TRAIN	TEST
Logistic Regression	0.59	0.5346534	0.4575	0.4752475
kNN	0.575	0.4059406	0.5425	0.4950495
Naïve Bayes	0.365	0.34653466	0.275	0.2673267
Decision Trees	1	0.7326732	1	1
Random Forest	1	0.7029703	1	1
Boosting	1	0.851485	1	1

and the ESG risk score is the least when controversy score is negligible.

Comparison of accuracies attained in determining the fraud and the risk levels by using different ML algorithms, is briefly illustrated in Table I. On comparing the accuracies attained by different ML models for risk analysis, Boosting gives the highest accuracy(i.e. 85.1485 %) among the other algorithms, while Naive Bayes gives the least accuracy(i.e. 34.653466 %). For fraud detction, Decision Trees, Random forest and Boosting, all three of them give the highest accuracies (100 % accuracies), while Naive Bayes gives the least accuracy(26.73267 %) in fraud detection as well.

ML algorithms in increasing order of their accuracies attained in risk analysis:

Naive Bayes $<$ kNN $<$ Logistic Regression $<$ Random Forest $<$ Decision Trees $<$ Boosting

ML algorithms in increasing order of their accuracies attained in fraud detection:

Naive Bayes $<$ Logistic Regression $<$ kNN $<$ $\{$Decision Trees, Random Forest, Boosting$\}$

VI. CONCLUSION AND FUTURE WORK

In this paper, a ML model for risk analysis and fraud detection in financial markets is designed and implemented using multiple algorithms. The results obtained show that the S&P 500 data set was successfully cleaned, analyzed and was able to give us desired results with high accuracies as depicted using Table I. In conclusion, it can be said confidentially that the research unlike the ones already done was able to do both risk analysis and fraud detection with the data set and that too with a great accuracy [2]. This makes the project unique, effective and powerful.

The research work is performed only on the S&P dataset , in the future it can be extended to various other data sets of the financial markets so that it can make a greater impact and can be useful for the common citizen to know the frauds and risks associated with different companies before buying and selling stocks. Furthermore, a mobile application or website can be designed using various techniques which uses our model to give the customers a pleasant and smooth experience. This will serve as an great asset in the financial markets as well as for the common people.

REFERENCES

[1] Analyzing the stock market based on the structure of knn network. *Chaos, Solitons Fractals*, 113:148–159, 2018.

[2] Studies from indiana university reveal new findings on financial management (the impact of competition and bank market regulation on banks' cost efficiency). pages 748–. NewsRX LLC, 2021.

[3] Stock market prediction based on statistical data using machine learning algorithms. *Journal of King Saud University - Science*, 34(4):101940, 2022.

[4] Retracted: Forecasting method of stock market volatility based on multidimensional data fusion. volume 2023, pages 1–1, 2023.

[5] Pritish Dugar. Sp 500 esg risk ratings, 2023.

[6] Maximilian Glück, Benjamin Hübel, and Hendrik Scholz. Esg rating events and stock market reactions. *Available at SSRN 3803254*, 2021.

[7] Manjusha Goel, Parveen Tomar, Priya Vinjamuri, G. Reddy, Mustafa Al-Taee, and Malik Alazzam. Using ai for predictive analytics in financial management. pages 963–967, 05 2023.

[8] Deepak Kumar, Pradeepta Sarangi, and Rajit Verma. A systematic review of stock market prediction using machine learning and statistical techniques. 49, 01 2021.

[9] Sujay Lokesh, Siddharth Mitta, Shlok Sethia, Srivardhan Reddy Kalli, and Manisha Sudhir. Risk analysis and prediction of the stock market using machine learning and nlp. volume 13, pages 16036–16041, 2018.

[10] Yilin Ma, Ruizhu Han, and Xiaoling Fu. Stock prediction based on random forest and lstm neural network. In *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, pages 126–130, 2019.

[11] Akib Mashrur, Wei Luo, Nayyar Zaidi, and Antonio Robles-Kelly. Machine learning for financial risk management: A survey. volume 8, pages 203203–203223, 01 2020.

[12] Anurag Nagar and Michael Hahsler. Using text and data mining techniques to extract stock market sentiment from live news streams. 02 2015.

[13] Kunal Pahwa and Neha Agarwal. Stock market analysis using supervised machine learning. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 197–200, 2019.

[14] Susmita Ray. A quick review of machine learning algorithms. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pages 35–39, 2019.

[15] Aleksandra Rzeźnik, Kathleen Weiss Hanley, and Loriana Pelizzon. The salience of esg ratings for stock pricing: Evidence from (potentially) confused investors. 2021.

[16] Faraz Sasani, Ramin Mousa, Ali Karkehabadi, Samin Dehbashi, and Ali Mohammadi. Tm-vector: A novel forecasting approach for market stock movement with a rich representation of twitter and market data. 2023.

[17] Navdeep Singh, Bharti Ramola, G. Reddy, Mustafa Al-Taee, and Malik Alazzam. The use of deep learning for predictive analytics in financial. pages 286–290, 05 2023.

[18] Sukhman Singh, Tarun Kumar Madan, Jitendra Kumar, and Ashutosh Kumar Singh. Stock market forecasting using machine learning: Today and tomorrow. In *2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*, volume 1, pages 738–745, 2019.

[19] Mrinalini Smita. Logistic regression model for predicting performance of sp bse30 company using ibm spss. 2021.

[20] Xiaoyuan Wan, Sichao Shen, and Jiachen Zhang. Do stock prices follow random walk over day and night? -– evidence from chinese stock market. pages 1–4, 2023.

[21] Huacheng Wang, Yanxia Jiang, and Hui Wang. Stock return prediction based on bagging-decision tree. In *2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009)*, pages 1575–1580, 2009.