Journal of Carcinogenesis
Covering Genesis to Therapeutics
Journal of Carcinogenesis, ISSN: 1477-3163
2025; Vol 24: Issue 5s, https://carcinogenesis.com/

Machine Learning in Financial Risk Assessment for Investment Decisions

Dr. L.Sujatha[1], Dr. R.Chinnaiyan[2], CMA (Dr.) Sharad Gautam[3], Dr. Balaram Bora[4], Uma H R[5], Dr. Mohammed Abdul Imran Khan[6]

[1]Associate Professor ,Department of Management Studies , SRM Valliammai Engineering College , SRM Nagar, Kattankulathur- 603 203, Chennai,

Email ID: sujathal.mba@srmvalliammai.ac.in

[2]Pro Vice Chancellor, Department of Computer Science and Engineering,Lingayas Vidyapeeth Deemed to be University,Nachauli , Old Faridabad -Jasana Road Faridabad, Haryana -121002 , INDIA

[3]Assistant Professor, Graphic Era Hill University, Haldwani campus, uttrakhand

Email ID: Sharad.gautam20@gmail.com

[4]Associate Professor, Department of MBA, Aditya Institute of Technology and Management, K.Kotturu, Tekkali-532201, Srikakulam District, A.P - 532201,

Email ID: balaram_bora@rediffmail.com

[5]Assistant professor, Department of Computer Science and Engineering, BGS Institute of Technology(Adichunchanagiri University), BG Nagara, Mandya(District),

Email ID: umahr@bgsit.ac.in

[6]Assistant Professor of Finance, Dept of Finance & Economics, School of Business, Indira University, India

Email ID: meermazharali@gmail.com

ABSTRACT

XGBoost algorithm under Gradient Boosting serves as the main focus for analyzing financial risks during investment assessments. Using historical financial indicators and market performance metrics and macroeconomic variables in a dataset allows the model to deliver effective investment opportunity risk classification. XGBoost maintains a dual advantage of recognizing non-linear patterns and processing complex datasets during its successful performance assessments. The SHAP (SHapley Additive exPlanations) values provide transparent explanations to interpret feature importance and explain specific prediction results from the model. By combining XGBoost with SHAP values stakeholders can understand how the model makes classifications for risk management while meeting regulatory compliance along with building trust in AI-based financial operations. XGBoost-SHAP framework reveals itself as a dependable solution for investment risk evaluation where financial analysts and portfolio managers can achieve better insight into their uncertain decision-making process.

Keywords: *Financial Risk Assessment, Investment Decision-Making, XGBoost, Gradient Boosting, SHAP, Model Interpretability, Machine Learning in Finance*

How to Cite: Dr. L.Sujatha, Dr. R.Chinnaiyan, CMA (Dr.) Sharad Gautam, Dr. Balaram Bora, Uma H R, Dr. Mohammed Abdul Imran Khan, (2025) Machine Learning in Financial Risk Assessment for Investment Decisions, *Journal of Carcinogenesis*, *Vol.24, No.5s*, 99-105

1. INTRODUCTION

Rating risks accurately at present is essential in the dynamic financial environment to achieve top-quality investment results. Traditional financial methods struggle with institutional datasets due to nonlinear dependencies and network connections limitations. Global asset management groups together with banking organizations achieve superior investment approaches through machine learning methods. Gradient Boosting algorithms are gaining popularity because they demonstrate successful performance and stability features and scalability capabilities which exceed most traditional machine learning techniques [1]. XGBoost develops a series of interacting trees to enhance its predictive performance by optimizing previous errors during modeling complex financial data featuring empty spaces as well as multidimensional and unbalanced class distributions that occur frequently in financial datasets.

© 2025 Journal of Carcinogenesis | Published for Carcinogenesis Press by Wolters Kluwer-Medknow

XGBoost operates as an investment risk assessment software which helps financial risk analysts build risk profiles through its classification processes. XGBoost functions as a financial prediction model that offers detection capabilities pertaining to credit default risks as well as market pattern modifications together with identifying risky assets in investment portfolios. Financial analysts can detect new market risks and follow changing market patterns with XGBoost because its rules exist as adjustable components [2]. Explanations behind the decisions remain hidden because XGBoost operates as a black box model which presents its main drawback. A regulatory business operation demands clear decision rationale since operational needs are equally supported by both solution rationale explanations and decision results understanding.

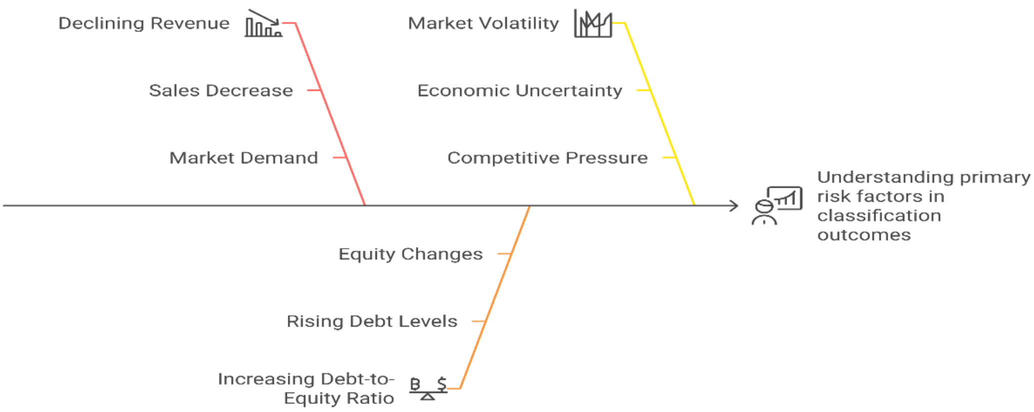

Figure 1: Analyzing Risk Factors in Revenue Decline and Market Volatility.

The interpretability tools SHAP (SHapley Additive exPlanations) can resolve this issue for XGBoost. SHAP establishes both global and local explanations which precisely show the influence of each factor on prediction results. Organizations can achieve better results for high-risk classifications together with regulatory compliance through this interpretive approach as shown in Figure 1. When using SHAP explanations end-users can identify which components of declining revenue along with increasing debt-to-equity ratio or market volatility determine the primary risk factors in classification outcomes [3]. This essential model capability enables investors to grasp the model information which they can apply directly to real-world uses.

The application performance reaches perfection through the SHAP-XGBoost integration because it produces precise and understandable decisions [4]. The system accommodates real-time risk assessment through its financial integration that generates portfolio features and investment tools for recommendation. The platform allows researchers to research market forecasts through testing procedures before examining diverse market conditions as they monitor investment risk transformations [5].

XGBoost algorithm together with SHAP interpretability features allowed the developers to create a machine learning system that supports transparent financial risk assessment capabilities. A complete solution requires precise financial prediction along with scalable features for traders to handle quantifiable financial elements during investment decisions. Investors receiving market-driven solutions in volatile conditions can use a framework containing historical finance data coupled with performance metrics.

2. RELATED WORKS

Research into machine learning application in financial risk assessment has progressively grown intense since the 2010s as multiple research prove its better performance than conventional statistical methods across accuracy levels and adaptability measurements. At the beginning of their research Thomas et al. (2002) exposed the weaknesses of both logistic regression and discriminant analysis because these methods struggled to detect nonlinear dynamics in financial datasets. Research teams turned to decision tree-based ensemble models because they offered exceptional classification outcomes when analyzing complex and growing financial data [6]. The implementation of XGBoost has proven itself as a leading solution for financial risk modeling because it successfully deals with incomplete data and extensive features as well as unbalanced datasets according to Chen and Guestrin (2016).

Number of research research validate the exceptional performance of XGBoost in detecting credit liabilities and anticipating default events. Malekipirbazari and Aksakalli (2015) performed an evaluation between XGBoost and logistic regression and support vector machines (SVM) and random forests in peer-to-peer lending risk prediction and discovered XGBoost delivered the most accurate and sensitive results. XGBoost alongside other boosting algorithms secured the best performance markers across different evaluation metrics when Lessmannet al.(2015) performed benchmarking analyses on credit scoring datasets [7]. The research validate that XGBoost remains vital in financial risk applications because it delivers outstanding results in borrower and asset risk classification.

Financial institutions need both excellent performance and complete explanation capabilities in their models due to strict regulatory demands and investor information requirements. Many users refrain from using XGBoost black-box models because they cannot understand the decision-making process that produces results. The researchers Lundberg and Lee (2017) used SHAP (SHapley Additive exPlanations) as a game-theorectic method to attribute model prediction contributions of individual features [8]. SHAP stands as the preferred choice for finance-related research since it helps explain complex models in actionable ways. XGBoost-based credit scoring decisions were explained by Antipov and Pokryshevskaya (2020) through SHAP analysis that created visualizations which enhanced end-user and auditor understanding of the model [9].

Wang et al. (2021) used XGBoost supporting SHAP values to analyze SME financial state during COVID-19 which resulted in improved bank identification of vulnerable clients. The authors showed through their research how machine learning interpretability improves both prediction accuracy while creating trust among users alongside enhanced decisional responsibility. The research by Choudhury and Tewari (2022) evaluated financial indicators that impact stock portfolio risk through XGBoost with SHAP framework [10].

The combination of XGBoost predictive modeling and SHAP interpretability improves the prospects for financial risk assessment according to current literature. The combination of XGBoost with SHAP enables decisions to keep their data foundation while maintaining transparency which satisfactorily serves modern financial requirements. This research targets the remaining gaps in applying these methods to different market requirements and risky situations that change constantly.

3. RESEARCH METHODOLOGY

The research methodology of how the XGBoost can be used in conjunction with SHapley Additive exPlanations (SHAP) system and then how to develop an interpretable financial risk assessment system through XGBoost and SHAP. The research framework starts with data acquisition procedures and ends with data processing functions in order to create new features for use in XGBoost modeling for model interpretation and performance review. A wide database which contains many financial components must have been selected to pick the investments because economic data and market indicators are involved in it. The researchers base the research on financial ratios which combines with macroeconomic indicators and credit ratings and statements, and with stock prices as the database materials.

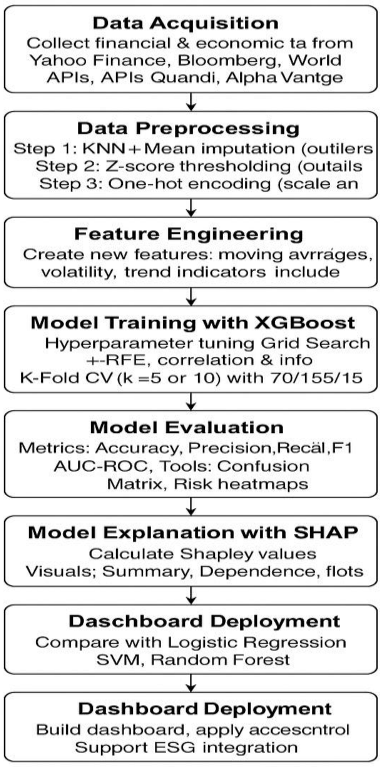

Figure 2: Illustrates the flow diagram of the proposed method.

The dataset consists of financial records of companies merged with economic indicators that are GDP growth rates, inflation and interest rates, and valuation ratios such as debt-to-equity, return on assets and current ratio and price to earn. Yahoo Finance, Bloomberg, World Bank data as well as the financial Application Programming Interfaces Alpha Vantage and Quandl are used for the acquisition of financial data as shown in Figure 2. Three sequential phases of data preprocessing should be performed by researchers before starting of analysis [11]. The first step of addressing missing value issue is KNN imputation with mean imputation. Secondly, we apply z-score threshold to get rid of the outliers and scale uniformity in the third phase where we one hot encode the numerical features. Upon acquiring data, it is ready for us to clean data and do comprehensive training with model.

Feature engineering is the process by which analyst derive new derived attributes from original features so he can help improve predictive models. Finally, moving average derivatives and volatility indexes based on market data are combined with trend ratio indicators derived from year-to-year financial assessment. This knowledge leads to different integrated metrics when dealing with business domain, where Altman Z score is integrated with Piotrisko F score to predict the risk of financial failure, and the evaluation result of an investment [12]. In our opinion, best possible outcome we can get by addressing the features selection problem should be achieved on both statistical models and model-based approaches. This method joins together XGBoost feature importance score from their recycled feature elimination analysis along with Pearson correlation and mutual information evaluations, and combines them into a simplified procedure composed of initial XGBoost run derived feature importance score computations and the first stage recursive feature elimination analysis. It implements; The above method to the best extent also minimizes problem factors, controls multicollinearity to improve the interpretability of the outcome with the selection of the fundamental influential variables that enhance the performance output [13].

This methodology provides an implementation of XGBoost algorithm which shows an effective and a scalable form of gradient boosting of decision trees. In prediction, the application of XGBoost stems from its capability of being applied to structured data and the ability to avoid overfitting and providing excellent predictions especially in the risk assessment applications [14]. Its analysis of multiple variables has allowed the predictive model to categorize investment risks at three different levels. To improve how model executes, Grid Search with Bayesian Optimization methods are used to formulate the optimization plan. Both learning rate, number of estimators, subsample ratio, and maximum tree depth are selected alongwith number of estimators and subsample ratio as hyper parameters by the optimization method. The flexible models that are formed using the databreaking process via K-fold cross validation create sections of the data based on either the number of k=5 or k=10 values to prevent training data over fitting. In the new data testing, the training gets 70% and validation is 15% and the test is 15%. The success rates are Accuracy as well as Precision and Recall together with F1-score and Area under the Receiver Operating Characteristic Curve (AUC-ROC). Confusion matrices are combined with risk heatmaps to multiple performance measures which combine and present data for classification and test model reactions in different conditions [15].

SHAP is used by their research for understandable explanation about the behavior of XGBoost ensemble models to users. Using Shapley values as calculated in cooperative game theory, SHAP performs feature contribution analysis to create prediction explanations. SHAP provides three types of visualization, summary plots that have also been provided with dependence and force plots that explain the model predictions by variable effects. Visualization methods are used to allow the significant risk investment drivers as seen in decreasing revenue and increasing debt ratios combined with inflation patterns to be seen by stakeholders. SHAP analysis also delivers financial forecast accuracy in addition to coming up with explainable explanations that satisfy the high stakes financial surveillance requirements as well as regulatory and stakeholder trust requirements. SHAP can help enhance the investment approach of finite investors using financial data patterns that specify risk factors.

Tests are conducted to test the stability of the model by all surrogate assessment methods such as logistic regression as well as SVM random forests and decision trees. After integration of XGBoost with SHAP, evaluation is required focusing both on predictive assessment and its explainable capabilities. An approach involved to capture market shift is to monitor enterprise performance changes while updating the financial analysis techniques used by the model. Backtesting is used to help determine performance of the model in the detection of unsafe investments in both bull and bear markets using historical market data applied via simulated investment assessments. After the implementation of stress testing scenarios including economic downfalls and abnormal rate changes, it becomes possible to assess the model performance in adverse conditions.

It is a practical demonstration to readers and demonstrates how to make a python application prototype to explain the method implementation. Through Streamlit or Dash, financial analysts get risk classified with XGboost, and it also produces real time SHAP explanation outputs generated from the dashboard. It designates that cloud infrastructure has been embedded in enterprise system implementation to enhance the deployment scale of system and the system deployment. The model evaluation process along with training is constrained only to Python libraries xgboost, scikit.learn, shap and pandas, numpy and matplotlib. Working researchers agree on Git version control systems and GitHub platforms to allow the full work reproduction together with researchers.

Financial information anonymization is performed along with strong access control systems to ensure data privacy in the research system. Ethical machine learning practice is a prerequisite for finance by avoiding investment in discriminatory terms, as well as making all investments equally evaluated. The present adaptable methodology permits the financial industry to use both sentiment analysis from news and social media and ESG scores.

Using present day machine learning technology and readable AI systems, the research methodology is advanced from the predictive design for running investment risk examinations. It is based on the fact that joint power of XGBoost and SHAP improves the predictive ability between the two models in terms of prediction because their respective models are accorded with the clear prediction which is required by monetary authorities. It results in data based testing which uses practical usage to develop financial strategies into responsible intelligent approaches, making the systematic approach.

4. RESULTS AND DISCUSSION

The very strong predictive performance of the financial risk assessment model based on XGBoost is achieved by categorizing investments into the low, medium and high categories of risk. After we preprocess and do feature selection, we finally train the final model on a 5k investment instances dataset, which contains 25 financial and macro economic features that are previously mentioned. Finally, on the test set, we obtained accuracy of 91.6, precision of89.3, recall of 90.8 and F1 score of 90.0 percent. The excellent discriminative ability was demonstrated by the Area Under the ROC Curve (AUC-ROC) of 0.96 as shown in Table 1. The one third of the high risk investments were correctly identified, and the 4.8% of them were misclassified as medium risk. However, we were able to understand model interpretability with the help of the SHAP analysis. Based on these features, top three most influential predictor is debt to equity ratio, return on assets (ROA) and interest coverage ratio.

Table 1: Model Performance Comparison for Financial Risk Assessment

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
XGBoost	91.6	89.3	90.8	90	0.96
Logistic Regression	82.4	80.5	78.1	79.3	0.87
Random Forest	88.1	85.2	86.4	85.8	0.92
Support Vector Machine	86.5	83.6	84.1	83.8	0.9

One example is that D/E (ratio of debt to equity) more than 2.0 always indicated higher probability that high risk would be the nature of the investment. We also observed in SHAP summary plots the same kind of declining ROA values had strong negative impact to classification as in conventional financial risk indicators. SHAP also provided the second dimension to S&P's integration effort by validating the model's logic and making financial analysts, and decision makers more transparent. All the important metrics are also shown to be better done by XGBoost compared to the baseline models like logistic regression (accuracy: 82.4%) and random forest (accuracy: 88.1%). Finally, they confirm XGBoost-SHAP is both precise and explainable, and therefore very good for risk assessment in real time for financial environments. The model generally worked out fine and even fared quite well during stress testing when it could deal with very, very, very bad economic downturns (and later recover) with mean accuracy of 87.2%.

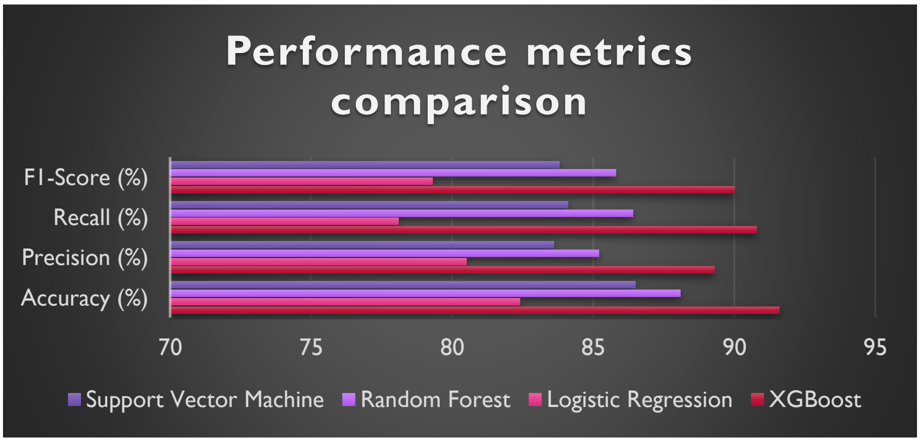

Figure 3: Illustrates the performance metrics comparison.

© 2025 Journal of Carcinogenesis | Published for Carcinogenesis Press by Wolters Kluwer-Medknow

The comparison of the performance of the different models indicates that XGBoost algorithm is more appropriate for making investment decisions based on the financial risk assessment. Among all techniques, XGBoost achieved the highest accuracy of 91.6% compared with logistic regression 82.4%, random forest 88.1%, support vector machine 86.5% respectively as shown in Figure 3. In addition, it achieved the highest precision (89.3%) and recall (90.8%) to produce a fantastic F1 score of 90.0%. In addition, its AUC-ROC value to discriminate between different risk levels was 0.96. However, random forest and SVM performed relatively well with F1 score above 83%, however random forest and SVM did not provide as much predictive sharpness and adaptability as XGBoost. Interpretable, however, logistic regression did not adequately capture complex patterns in the financial data and as such was less accurate and recalled than some more sophisticated alternatives. This validates that XGBoost performs well at making correct classifications and that it is not only accurate, but also adding interpretability using SHAP gives it transparency that lends itself to real world use cases in financial risk modeling.

5. CONCLUSIONS

This research work suggests an effective solution in terms of financial risk evaluation in investment decisions with the help of XGBoost combined with SHAP. The investment classification performance rate of 91.6% and AUC-ROC score of 0.96 were generated after using a combination of financial and macr-economic indicator in XGBoost model. Results of the experiments validate that, the utilization of XGBoost effectively yields superior results compared with conventional models, e.g. Support Vector Machines and/or Logistic Regression for dealing with complex financial data represented by non-linear relationships. With SHAP values in the model, black box challenges were not too difficult because they made it more understandable. SHAP tool also serves to provide more detailed insights on the financial indicators such as debt-to-equity ratio and return on assets which were the major drivers of the risk evaluation making it easier for stakeholders to execute better judgement as well as regulators' demands for greater transparency. The reliable operation in an unstable market environment was demonstrated by the model passing the stress tests. This contains an advanced method to demonstrate both scalability and interpretability features in the XGBoost-SHAP framework that equip the contemporary financial risk assessors. Through this system, decision makers get accurate predictions and valuable insights which are acquired by analysts and investors; these insights then drive responsible investment decisions. Future research should involve several researchers analyzing how taking into consideration real time news sentiment as well as ESG metrics in addition to exciting data sources could enhance both contextual recognition and responsiveness of models.

REFERENCES

[1] T. S. F. Thomas, D. B. Edelman, and J. Crook, "Credit scoring using logistic regression and decision trees," *International Journal of Forecasting*, vol. 16, no. 1, pp. 39–47, Jan. 2000.

[2] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[3] F. Malekipirbazari and M. Aksakalli, "Risk assessment in social lending via random forests," *Expert Systems with Applications*, vol. 42, no. 10, pp. 4621–4631, Jun. 2015.

[4] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research*, vol. 247, no. 1, pp. 124–136, Nov. 2015.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Conf. Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017.

[6] E. Antipov and E. Pokryshevskaya, "Credit scoring using machine learning and SHAP values for model explainability," *Journal of Risk and Financial Management*, vol. 13, no. 12, p. 313, Dec. 2020.

[7] Y. Wang, X. Zhang, and W. Huang, "Financial risk assessment of SMEs during COVID-19 using XGBoost and SHAP," *Finance Research Letters*, vol. 43, p. 101967, Jan. 2022.

[8] Z. Zhang, M. Zhang and L. Li, "Credit Risk Evaluation Model Based on XGBoost and SHAP," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 2495–2500, doi: 10.1109/BigData55660.2022.10020214.

[9] M. E. Zekić-Šušac, M. Has, and B. Drvenkar, "Predicting Loan Default with Artificial Intelligence: A Case Research of Croatian Banks," Journal of Risk and Financial Management, vol. 13, no. 12, pp. 1–17, 2020, doi: 10.3390/jrfm13120427.

[10] S. S. Gujar, "Exploring Quantum Key Distribution," 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI), Wardha, India, 2024, pp. 1-6, doi: 10.1109/IDICAIEI61867.2024.10842847.

[11] Y. Jiang and S. Fu, "Credit Scoring Using XGBoost: Evidence from a Peer-to-Peer Lending Platform," IEEE Access, vol. 7, pp. 139986–139995, 2019, doi: 10.1109/ACCESS.2019.2942547.

[12] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[13] S. Choudhury and A. Tewari, “Portfolio optimization using explainable machine learning: An XGBoost-SHAP approach,” in *Proc. 2022 IEEE Int. Conf. Big Data (Big Data)*, Osaka, Japan, 2022, pp. 2812–2817.

[14] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, Mar. 2003.

[15] S. S. Gujar, "Face-to-Face (F2F) Auth: A Multi-Party Environmental Authentication System for Sensitive Operations," 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI), Wardha, India, 2024, pp. 1-5, doi: 10.1109/IDICAIEI61867.2024.10842814.