PREDICTING BANKRUPTCY WITH MACHINE LEARNING MODELS

G Naga Sujini[1], Runku Yasaswini [2]

[1] Department of CSE, Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad, 500075, Telangana, India.

[2] Department of IT, Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad, 500075, Telangana, India.

---***---

Abstract - Bankruptcy prediction is a pivotal task in finance and accounting, offering stakeholders critical insights into a firm's solvency. This study explores multiple machine learning classifiers, including XGBoost, Random Forest, Decision Trees, Neural Networks, SVM, Naïve Bayes, and Logistic Regression, to predict corporate bankruptcy. Using the Polish bankruptcy dataset from UCI, various preprocessing techniques such as EM and K-NN imputation and SMOTE were applied to handle data imbalance and missing values. Extensive performance evaluations using cross-validation reveal that ensemble methods like XGBoost and Random Forest offer superior predictive power, providing a robust foundation for decision-making. The study also presents a web-based prediction tool developed using Flask for real-time assessment. Future research is directed towards integrating deep learning architectures and multi-source datasets for enhanced generalizability.

Key Words: Bankruptcy Prediction, Machine Learning, Financial Distress, XGBoost, Random Forest, Data, SMOTE.

1.INTRODUCTION

The financial stability of a company is critical for investors, creditors, and regulators. The timely prediction of bankruptcy can mitigate risks and avert major financial losses. Traditional methods often fall short in dynamic markets, where machine learning provides more adaptive solutions. With access to massive datasets and financial metrics, this domain is ideal for AI-driven analysis. The purpose of this study is to build, compare, and deploy several ML models to predict company bankruptcy with high accuracy.

1.1 Motivation

Corporate bankruptcies affect not just shareholders and employees but also investors, creditors, regulatory agencies, and the broader economy. Early detection of financial distress empowers stakeholders to take preemptive measures such as credit restructuring or operational pivoting. With the increasing digitization of financial records and the availability of open-source datasets, artificial intelligence (AI) and machine learning (ML) provide scalable solutions to this problem.

From a technical standpoint, bankruptcy prediction benefits from AI's capability to model complex non-linear relationships in financial data—relationships often overlooked by traditional econometrics. Companies maintain quarterly and annual reports that contain ratios on profitability, liquidity, leverage, and operational efficiency, making this domain a rich field for data mining and machine learning.

Moreover, recent financial crises have underscored the need for proactive monitoring tools. For example, the 2008 global financial meltdown could have been better mitigated with early-warning predictive systems at the firm-level. This motivates the use of ML models trained on historical data to forecast insolvency years in advance.

1.2 Problem Statement

Despite the advancements in methods and tools, bankruptcy prediction faces two primary challenges that hinder the development of effective models. First, while domain experts propose various econometric indicators to describe a firm's condition, there is no clear methodology for combining these indicators into a successful predictive model. The absence of standardized practices often results in models that lack consistency or fail to generalize across different datasets. Second, the training data used for predictive models are typically imbalanced, as the number of successful companies far outweighs the number of bankrupt ones. This imbalance leads to models that are biased towards predicting success (majority class), even in cases where a company is in financial distress, ultimately reducing the accuracy of predictions for the minority class.

1.3 Existing System

The current state-of-the-art integrates financial ratios with network-based machine learning. Authors like Kadkhoda and Amiri have utilized correlation-based company networks combined with classifiers like Random Forests and SVMs. These systems show promising accuracy (up to 94%) but suffer from key limitations:

- **Computational Overhead:** Network construction is computationally intensive and may not scale for real-time prediction.
- **Data Dependency:** These models are sensitive to missing or noisy financial entries, which are common in real-world datasets.
- **Static Design:** Many systems cannot accommodate live financial feeds or user inputs, making them less useful in applied settings.

1.4 Proposed System

The proposed bankruptcy prediction system leverages the UCI Polish bankruptcy dataset to predict the likelihood of company bankruptcies, utilizing key financial ratios such as profitability ratios, debt ratios, and liquidity metrics. This system aims to predict whether a company is at risk of bankruptcy based on its financial data, providing valuable insights to investors, creditors, and other stakeholders. The system focuses on assessing the financial stability of companies using machine learning models to classify them as either likely to face bankruptcy or not. The goal is to develop an efficient, accurate, and user-friendly tool for evaluating a company’s financial health.

The proposed system introduces several key advancements over the existing systems. One of the major improvements is in the data preprocessing phase, where advanced techniques like K-Nearest Neighbours (K-NN) and Expectation-Maximization (EM) imputation are employed to handle missing data more effectively, ensuring higher data quality and more reliable predictions. Additionally, the use of the Synthetic Minority Oversampling Technique (SMOTE) addresses class imbalance by generating synthetic examples of the minority class (companies at risk of bankruptcy), leading to better model performance. The system also evaluates a diverse set of machine learning models, including Logistic Regression, Decision Trees, Random Forest, XGBoost, Support Vector Machines (SVM), Naive Bayes, and Neural Networks.

2. LITERATURE SURVEY

[1] In the paper titled *"Predicting Financial Distress in High-Dimensional Imbalanced Datasets: A Multi-Heterogeneous Self-Paced Ensemble Learning Framework (MDPI-2025)"* by Ruize Gao, Shaoze Cui, Yu Wang, and Wei Xu, the authors address the complex challenge of financial distress prediction (FDP), which is hindered by high-dimensional features and data imbalance. They propose FinMHSPE, a novel ensemble learning framework that incorporates pairwise data comparisons across multiple timeframes and uses the maximum relevance minimum redundancy (mRMR) algorithm for efficient feature selection, enhancing model performance in predicting corporate solvency.

[2] In the paper titled *"A Multi-Stage Financial Distress Early Warning System: Analyzing Corporate Insolvency with Random Forest (Springer-2025)"* by Katsuyuki Tanaka, Takuo Higashide, Takuji Kinkyo, and Shigeyuki Hamori, the authors introduce an early warning system to differentiate between insolvency and bankruptcy. By applying a Random Forest classifier, they identify distinct financial indicators for each stage of distress and show that insolvency can act as an intermediate state with unique predictive variables.

[3] In the paper titled *"Business Failure Prediction Based on a Cost-Sensitive Extreme Gradient Boosting Machine (IEEE-2024)"* by Yao Zou, Changchun Gao, and Han Gao, the authors focus on overcoming data imbalance and poor interpretability in business failure prediction (BFP). Their proposed cost-sensitive XGBoost algorithm enhances predictive accuracy, outperforming traditional statistical techniques and earlier machine learning models by effectively handling minority class underrepresentation.

[4] In the paper titled *"A Hybrid Metaheuristic Method in Training Artificial Neural Network for Bankruptcy Prediction (IEEE-2023)"* by Abdollah Ansari, Ibrahim Said Ahmad, Azuraliza Abu Bakar, and Mohd Ridzwan Yaakub, the authors develop a hybrid optimization technique combining the Magnetic Optimization Algorithm (MOA) and Particle Swarm Optimization (PSO) to improve the training efficiency of Artificial Neural Networks (ANNs). Their hybrid MOA-PSO model achieves 99.7% accuracy in bankruptcy prediction, showing superior performance over other optimization algorithms.

[5] In the paper titled *“An Intelligent Approach for Predicting Bankruptcy Empowered with Machine Learning Technique (IEEE-2022)”* by Vivek Kothuru et al., the authors perform a comparative study of advanced machine learning models such as Random Forest, XGBoost, and SVM. Their analysis demonstrates that ensemble learning methods provide superior accuracy and robustness for bankruptcy prediction, supporting improved financial decision-making.

[6] In the paper titled *“Predicting Company Bankruptcy Using Random Forest Method (IEEE-2021)”* by Maria Susan Anggreainy, Vincent, Ishika Gurnani, and Febryan Stefanus Tandian, the authors utilize a Random Forest approach to analyze a dataset of over 6,800 firms from 1999 to 2009. Achieving 97.8% accuracy, their findings validate Random Forest as a highly effective tool for predicting bankruptcy in large corporate datasets.

[7] In the paper titled *“Utilizing Principal Component Analysis to Enhance Machine Learning in Bankruptcy Prediction: A Comparative Investigation (Proc. 6th Int. Conf. Computer. Data Sci., 2024)”* by Xinming Ma et al., the authors explore how Principal Component Analysis (PCA) can optimize machine learning models in the context of bankruptcy forecasting. Using a Taiwanese financial dataset, they demonstrate that PCA reduces dimensionality while improving logistic regression model accuracy, suggesting PCA's effectiveness in preprocessing financial data.

3. DESIGN AND METHODOLOGY

3.1 System Architecture

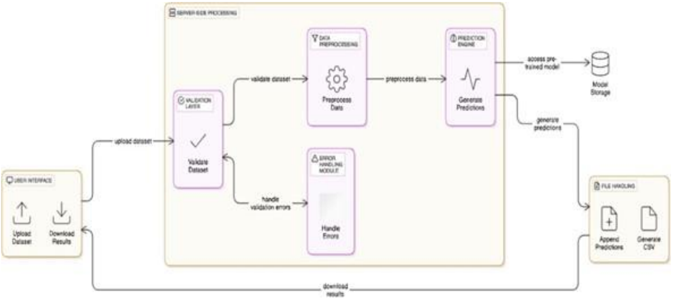


Fig -1: System Architecture Diagram

The system architecture, illustrated in Fig -1, consists of three key modules: User Interface, Server-Side Processing, and File Handling. The User Interface allows users to upload financial datasets (CSV) and download predictions, ensuring a smooth, dependency-free user experience. The Server-Side Processing module handles file validation, data preprocessing, and prediction using a pre-trained machine learning model. It also includes error handling for robust operation. The File Handling module appends predictions to the original dataset and generates a downloadable CSV file.

3.2 Dataset

The UCI Polish bankruptcy dataset was selected due to its rich 64-feature set and 5-year bankruptcy indicators.

3.3 Data Preprocessing

Imputation techniques:

- **Mean Imputation:** Simple average replacement.
- **K-NN Imputation:** Contextual filling based on nearest neighbors.
- **EM Imputation:** Statistically driven iterative estimation.

To address class imbalance, **SMOTE** was used for oversampling the minority class (bankrupt).

3.4 Model Training and Selection

Models used:

- Logistic Regression
- Gaussian Naïve Bayes
- Decision Trees
- Random Forests
- XGBoost
- Neural Networks
- Support Vector Machine

Each model underwent 5-fold cross-validation, with metrics computed: Accuracy, Precision, Recall, AUC.

3.5 Hyperparameter Tuning

- XGBoost was tuned using learning rate (0.1), max depth (5), and 100 estimators.
- Neural networks used a 3-layer feedforward structure with relu activations.

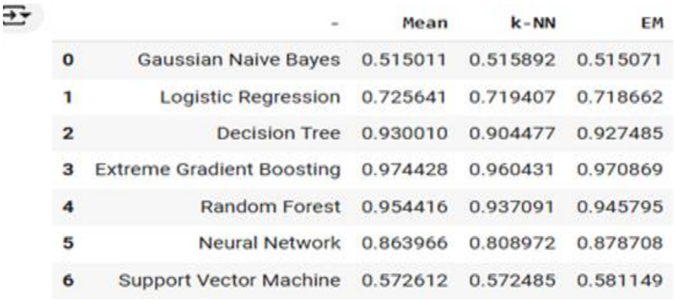
4. IMPLEMENTATION AND DEPLOYMENT

The system was built using:

- **Flask**: Backend server for model serving.
- **Pandas, NumPy, Sklearn, Joblib**: For data processing and model loading.
- **Joblib**: For serialization of the final trained XGBoost model.

4.1 Model Summary

presents a comparative analysis of different machine learning algorithms based on their performance metrics.

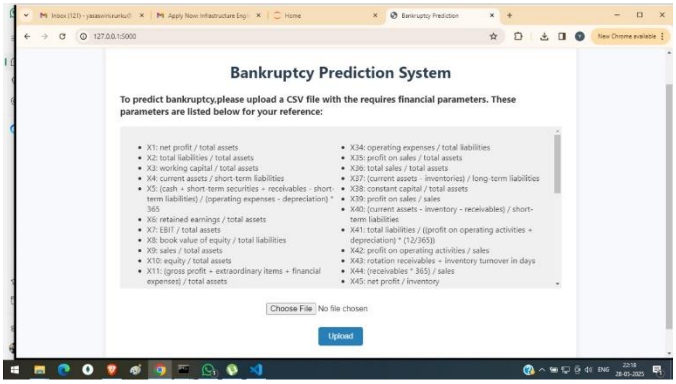


	-	Mean	k-NN	EM
0	Gaussian Naive Bayes	0.515011	0.515892	0.515071
1	Logistic Regression	0.725641	0.719407	0.718662
2	Decision Tree	0.930010	0.904477	0.927485
3	Extreme Gradient Boosting	0.974428	0.960431	0.970869
4	Random Forest	0.954416	0.937091	0.945795
5	Neural Network	0.863966	0.808972	0.878708
6	Support Vector Machine	0.572612	0.572485	0.581149

Fig -2: Model Accuracy Metrics Using Mean, k-NN, and EM

4.2 Results

In the results, each company is assigned a binary label under the "BM" column, where the value **0** signifies that the company is financially stable and not at risk of bankruptcy, while the value **1** indicates a high likelihood of the company going bankrupt based on its financial attributes. These classifications are derived from the machine learning model's analysis of historical financial indicators, allowing users to easily interpret the financial status of each entity within the dataset.

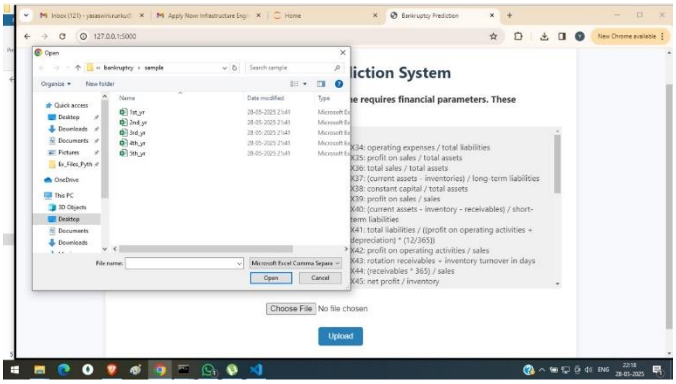


Bankruptcy Prediction System

To predict bankruptcy,please upload a CSV file with the requires financial parameters. These parameters are listed below for your reference:

- X1: net profit / total assets
- X2: total liabilities / total assets
- X3: working capital / total assets
- X4: current assets / short-term liabilities
- X5: (cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation) * 365
- X6: retained earnings / total assets
- X7: EBIT / total assets
- X8: book value of equity / total liabilities
- X9: sales / total assets
- X10: equity / total assets
- X11: (gross profit + extraordinary items + financial expenses) / total assets
- X34: operating expenses / total liabilities
- X35: profit on sales / total assets
- X36: total sales / total assets
- X37: (current assets - inventories) / long-term liabilities
- X38: constant capital / total assets
- X39: profit on sales / sales
- X40: (current assets - inventory - receivables) / short-term liabilities
- X41: total liabilities / ((profit on operating activities + depreciation) * (12/365))
- X42: profit on operating activities / sales
- X43: rotation receivables + inventory turnover in days
- X44: (receivables * 365) / sales
- X45: net profit / inventory

Fig -3: Homepage Interface



Name	Date modified	Type
1st_yr	28-05-2025 21:41	Microsoft Ex
2nd_yr	28-05-2025 21:41	Microsoft Ex
3rd_yr	28-05-2025 21:41	Microsoft Ex
4th_yr	28-05-2025 21:41	Microsoft Ex
5th_yr	28-05-2025 21:41	Microsoft Ex

iction System

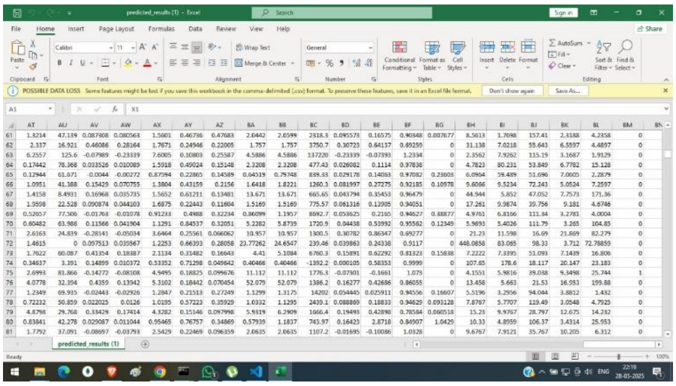
e requires financial parameters. These

X34: operating expenses / total liabilities
X35: profit on sales / total assets
X36: total sales / total assets
X37: (current assets - inventories) / long-term liabilities
X38: constant capital / total assets
X39: profit on sales / sales
X40: (current assets - inventory - receivables) / short-term liabilities
X41: total liabilities / ((profit on operating activities + depreciation) * (12/365))
X42: profit on operating activities / sales
X43: rotation receivables + inventory turnover in days
X44: (receivables * 365) / sales
X45: net profit / inventory

Choose File No file chosen

Upload

Fig -4: Uploading CSV File



POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

A1 X1

	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM
61	1.3214	47.139	0.087308	0.080563	1.5601	0.46736	0.47683	2.0442	2.0599	2318.3	0.095573	0.16575	0.90348	0.007677	8.5613	1.7698	157.41	2.3188	4.2358	0
62	2.337	16.921	0.46086	0.28164	1.7671	0.24946	0.22005	1.757	1.757	3750.7	0.30723	0.64137	0.69259	0	31.138	7.0218	55.643	6.5597	4.4897	0
63	6.2557	125.6	-0.07989	-0.23339	7.6005	0.10803	0.25587	4.5886	4.5886	137220	-0.23339	-0.07393	1.2334	0	2.3562	7.9262	115.19	3.1687	1.9129	0
64	0.17442	78.368	0.033526	0.010089	1.5918	0.49024	0.15148	2.3208	2.3208	477.43	0.026082	0.1114	0.97838	0	4.7823	80.231	53.849	6.7782	15.128	0
65	0.12944	61.671	-0.0044	-0.00272	0.87594	0.22865	0.14589	0.64519	0.79748	839.33	0.029178	0.14063	0.97082	0.23603	6.0964	59.489	51.696	7.0605	2.2879	0
66	1.0951	41.388	0.15429	0.070755	1.3804	0.43159	0.2156	1.6418	1.8221	1260.3	0.081997	0.27275	0.92185	0.10978	9.6066	9.5234	72.243	5.0524	7.2597	0
67	1.4158	8.4931	0.16968	0.035735	1.5652	0.61211	0.13481	13.671	13.671	665.65	0.043794	0.35453	0.96479	0	44.944	5.852	47.052	7.7573	171.36	0
68	1.9598	22.528	0.090874	0.044103	1.6875	0.22443	0.11604	1.5169	1.5169	775.57	0.061316	0.13905	0.94051	0	17.261	9.9874	39.756	9.181	4.6746	0
69	0.52657	77.506	-0.01763	-0.01078	0.91233	0.4988	0.32234	0.86099	1.1957	8692.7	0.053625	0.2165	0.94627	0.38877	4.9761	6.8166	111.34	3.2781	4.0004	0
70	0.60482	63.986	0.11566	0.041904	1.1291	0.84537	0.32051	5.2282	5.8739	1720.9	0.04438	0.53992	0.95562	0.12349	5.9693	5.4026	111.79	3.265	104.85	0
71	2.6163	24.839	-0.28141	-0.05034	3.6464	0.25561	0.066062	10.957	10.957	1300.5	0.30782	0.86347	0.69277	0	21.23	11.598	16.69	21.869	82.279	0
72	1.4615	0	0.097513	0.039567	1.2253	0.66393	0.28058	23.77262	24.6547	239.46	0.039863	0.24338	0.9117	0	448.0858	83.065	98.33	3.712	72.78859	0
73	1.7622	60.087	0.41354	0.18387	2.1134	0.31482	0.16643	4.41	5.1084	6760.3	0.15891	0.62292	0.81323	0.15838	7.2222	7.3395	51.093	7.1439	16.806	0
74	0.34637	3.391	0.14899	0.010372	0.53352	0.71298	0.049642	0.40466	0.40466	-1392.2	0.000105	0.58353	0.9999	0	107.65	178.6	18.117	20.147	23.183	0
75	2.6993	81.866	-0.14272	-0.08108	4.9495	0.18825	0.099676	11.112	11.112	1776.3	-0.07301	-0.1661	1.073	0	4.1551	5.9816	39.038	9.3498	25.744	1
76	4.0778	32.394	0.4359	0.13942	5.3102	0.18442	0.070454	52.079	52.079	1386.2	0.16277	0.42686	0.86055	0	13.458	5.663	21.53	16.953	199.88	0
77	1.2349	69.935	-0.02443	-0.02926	1.2847	0.21513	0.27249	1.1299	1.3175	14202	0.054445	0.025911	0.94556	0.16607	5.5196	3.2956	94.044	3.8812	1.432	0
78	0.72232	50.859	0.022025	0.0126	1.0195	0.57223	0.35929	1.0332	1.1295	2439.1	0.088869	0.18833	0.94629	0.093128	7.8767	5.7707	119.49	3.0548	4.7925	0
79	4.8798	29.768	0.33429	0.17414	4.3282	0.15146	0.097998	5.9319	6.2909	1666.4	0.19493	0.42898	0.78584	0.060518	15.23	9.9767	28.797	12.675	14.232	0
80	0.83841	42.278	0.029087	0.011044	0.95465	0.76757	0.34869	0.57939	1.1837	743.97	0.16423	2.8718	0.84907	1.0429	10.33	4.8959	106.37	3.4314	25.953	0
81	1.7792	37.091	-0.08697	-0.03793	2.5429	0.22469	0.096359	2.0635	2.0635	1107.2	-0.01695	-0.10086	1.0328	0	9.6767	7.9121	35.767	10.205	6.312	0

Fig -5: predicted result

5. CONCLUSION

Bankruptcy prediction is a critical task that plays a significant role in finance and accounting by helping creditors, investors, and stakeholders assess the financial health of companies. In this study, we explored various machine learning techniques, such as Extreme Gradient Boosting, Random Forests, Support Vector Machines, and Neural Networks, among others, to predict bankruptcy. By leveraging a Polish bankruptcy dataset, we incorporated synthetic features derived from econometric measures to enhance the predictive power of these models. The data preprocessing phase addressed key challenges, including missing data and class imbalance, using techniques like imputation (Mean, k-Nearest Neighbours, MICE) and SMOTE oversampling. Model evaluation through metrics such as accuracy, precision, recall, and cross-validation provided valuable insights into the performance of these models.

5.1 FUTURE SCOPE

While the study successfully demonstrated the application of machine learning models to bankruptcy prediction, there remains room for improvement and further exploration. Future research could focus on incorporating additional datasets from diverse geographical regions to improve model generalizability across industries and economies. Experimenting with more advanced techniques, such as ensemble learning and deep learning architectures, may further enhance prediction accuracy. Additionally, explainability in machine learning models could be emphasized, enabling stakeholders to interpret the results more effectively. Addressing evolving financial trends and integrating real-time data could also make the models more adaptive to market dynamics.

REFERENCES

[1] Gao, R., Cui, S., Wang, Y. et al. "Predicting financial distress in high-dimensional imbalanced datasets: a multi-heterogeneous self-paced ensemble learning framework." Financ Innov 11, 50 (2025). https://doi.org/10.1186/s40854-024-00745-w.

[2] Tanaka, K.; Higashide, T.; Kinkyo, T.; Hamori, S. "A Multi-Stage Financial Distress Early Warning System: Analyzing Corporate Insolvency with Random Forest." *J. Risk Financial Manag.* 2025, *18*, 195.

[3] Y. Zou, C. Gao and H. Gao, "Business Failure Prediction Based on a Cost-Sensitive Extreme Gradient Boosting Machine," in *IEEE Access*, vol. 10, pp. 42623-42639, 2022, doi: 10.1109/ACCESS.2022.3168857.

[4] A. Ansari, I. S. Ahmad, A. A. Bakar and M. R. Yaakub, "A Hybrid Metaheuristic Method in Training Artificial Neural Network for Bankruptcy Prediction," in *IEEE Access*, vol. 8, pp. 176640-176650, 2020,

[5] N. e. Radwan, H. M. Alzoubi, N. Sahawneh, A. Fatima, A. Rehman and S. Khan, "An Intelligent Approach for Predicting Bankruptcy Empowered with Machine Learning Technique," *2022 International Conference on Cyber Resilience (ICCR)*, Dubai, United Arab Emirates, 2022, pp. 1-5, doi: 10.1109/ICCR56254.2022.9995890.

[6] I. Gurnani, Vincent, F. S. Tandian and M. S. Anggreainy, "Predicting Company Bankruptcy Using Random Forest Method," *2021 2nd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, IPOH, Malaysia, 2021, pp. 1-5, doi: 10.1109/AiDAS53897.2021.9574384.

[7] Ma, X. (2024). "Utilizing principal component analysis to enhance machine learning in bankruptcy prediction: A comparative investigation". *Applied and Computational Engineering*, *68*(1), 306–312. https://doi.org/10.54254/2755-2721/68/20241500.