A Machine Learning Based Framework For Bankruptcy Prediction In Corporate Finances Using Explainable AI Techniques

Tabassum Maktum[1], Namita Pulgam[2], Vaishnavi Chandgadkar[3], Prachi Pathak[3] and Aditya Solanki[3]
[1]Department of Computer Engineering, School of Engineering & Technology, Anjuman I Islam's Kalsekar Technical Campus, New Panvel, Maharashtra, India.
[2]Department of Computer Science and Engineering (Data Science), D J Sanghvi College of Engineering, Mumbai, Maharashtra, India
[3]Department of Computer Engineering, Ramrao Adik Institute of Technology, D Y Patil Deemed to be University, Navi Mumbai, Maharashtra, India.
tabassum.maktum@aiktc.ac.in, Namita.Pulgam@djsce.ac.in, chandgadkar.vaishnavi@gmail.com, prachii1006@gmail.com, adityasolanki303@gmail.com,

Keywords: Bankruptcy, quick ratio, debt-equity ratio, Altman Z-Score, explainable AI, random forest

Received: July 20, 2024

Forecasting bankruptcy within corporate finances is an indispensable endeavor crucial for sustaining business growth and fostering stability. The paper presents a methodology to redefine the conventional approach to bankruptcy prediction within corporate finance. Through the adept utilization of advanced machine learning techniques, notably classification models, a dynamic and adaptable framework is established, enabling the systematic categorization of companies based on their bankruptcy risk profiles. Moreover, the methodology addresses the inherent challenge of data bias by integrating oversampling techniques like the Synthetic Minority Over-sampling Technique (SMOTE), thereby ensuring a more equitable representation of minority class samples and bolstering the model's predictive accuracy. The resulting model delivers timely and precise forecasts of bankruptcy risk, fortified by crucial recommendations such as the Altman Z-Score for vulnerability assessment, Debt-to-Equity Ratio for insights into leverage, Quick Ratio for assessing liquidity, and Explainable AI Techniques like SHapley Additive exPlanations (SHAP) analysis for transparent interpretations. This comprehensive approach equips stakeholders with tailored recommendations, empowering them to proactively safeguard their organizations' financial well-being and avert the perils of bankruptcy. The comparative analysis presented in paper demonstrates that the proposed method assesses the bankruptcy risk more accurately. The integration of Explainable AI techniques and key financial metrics helps the stakeholders to take vital decisions about corporate finances.

Povzetek: Opisano je napovedovanje stečaja v podjetniških financah z uporabo tehnik razložljive umetne inteligence. Z uporabo naprednih tehnik strojnega učenja, kot so klasifikacijski modeli in tehnike prekomernega vzorčenja (SMOTE), avtorji razvijejo dinamičen okvir za kategorizacijo podjetij glede na tveganje stečaja. Vključitev meritev, kot so Altmanov Z-izid, razmerje med dolgom in kapitalom ter hitro razmerje, skupaj z analizo SHAP, zagotavlja pregledne interpretacije in praktične vpoglede za zainteresirane strani.

1 Introduction

In the intricate landscape of corporate finance, the looming specter of bankruptcy poses a significant challenge, demanding agile manoeuvring to shield stakeholders' interests and ensure the sustained vitality of the organization. When a company teeters on the brink of insolvency, its reverberations are profound, affecting every facet of its operations, financial stability, and brand reputation. Bankruptcy, a stark manifestation of financial distress, not only inflicts substantial losses upon investors but also serves as a stark indicator of underlying operational inefficiencies, strategic miscalculations, or external economic pressures. In today's volatile and dynamic global marketplace, the proactive anticipation and adept management of bankruptcy risk have become imperative for modern enterprises striving for longevity and competitiveness. Amidst the prevailing uncertainties and cut-throat competition, the ability to anticipate and pre-emptively mitigate the threat of bankruptcy emerges as a pivotal strategic priority. Acknowledging the nuanced nature of financial distress, companies must employ robust analytical frameworks, sophisticated risk assessment methodologies, and well-orchestrated contingency plans to effectively identify vulnerabilities and fortify their financial resilience. By integrating bankruptcy anticipation as a cornerstone of their strategic approach, businesses not only fortify their resilience against unforeseen economic downturns but also foster sustainability and engender long-term value creation. This proactive stance not only shields against potential crises but also lays the groundwork for enduring success in an ever-evolving, interconnected global economy, positioning organizations to navigate turbulent waters with confidence and agility.

Traditional systems for bankruptcy prediction often lack explainability regarding how financial indicators contribute to predictions and remain static, failing to adapt to changing economic conditions. Moreover, they may suffer from imbalanced datasets, biased towards nonbankrupt instances, and rely on linear assumptions, potentially overlooking nonlinear patterns. This imbalance in the dataset, favouring non-bankrupt instances, may lead to an underestimation of bankruptcy risks. Therefore, there is a pressing demand for an innovative approach capable of accurately forecasting bankruptcy in corporate finances, empowering stakeholders with timely insights to proactively safeguard their organizations' financial stability. Recognizing the pivotal role of timely recognition of financial distress in ensuring sustained business growth and solidity, the proposed methodology aims to redefine bankruptcy prediction in corporate finance, ushering in an era of proactive risk management. Hence, proposed methodology sets a new standard for bankruptcy prediction, empowering businesses to navigate the complexities of the modern economic landscape with agility and confidence. Below objectives outline the key components of the proposed methodology aimed at addressing the challenges of bankruptcy prediction within corporate finance:

– Develop a predictive analytics framework for timely insights into financial vulnerabilities, ensuring stability.
– Implement a model using the Altman Z Score to assess financial health and bankruptcy risk.
– Use the Debt Equity Ratio to evaluate financial leverage and bankruptcy risk.
– Employ the Quick Ratio to gauge liquidity and signal potential short-term distress.
– Analyze SHAP values to interpret predictions, identify key bankruptcy factors, and offer actionable insights to mitigate risks.

The paper underscores the critical importance of forecasting bankruptcy in corporate finances, highlighting the need for early identification of financial instability and proactive risk mitigation. It reviews traditional and modern bankruptcy prediction methods, including machine learning and oversampling techniques to handle data bias. The proposed methodology focuses on advanced machine learning techniques and oversampling. Results are presented and evaluated, followed by a summary of key findings, implications for stakeholders, limitations, and suggestions for future research.

2 Related work

Evident research done in [1] showcases the efficacy of the Kmeans-SMOTE Integration method in addressing the challenge of imbalanced data within the context of classifying financial distress companies. This study yielded notable results. Support Vector Machine (SVM) achieved an impressive accuracy rate of 99.1%, while Na¨ıve Bayes demonstrated a respectable accuracy of 87%. The success of these models was attributed to the utilization of the Kmeans-SMOTE method, which effectively mitigated data imbalance. This approach not only enhanced classification accuracy but also underscored its potential to bolster the reliability of financial distress prediction systems.

Research in [2] demonstrated the effectiveness of machine learning and data balancing techniques for bankruptcy prediction. Bagging achieved 90.75% accuracy, with Random Forest (RF) at 91.61%. Support Vector Machine Linear (SVML) and Radial (SVMR) reached 80.06% and 79.96%, respectively, while Artificial Neural Network (ANN) and Decision Tree (DT) achieved 81.92% and 81.43%. The study utilized five data balancing methods—random oversampling, SMOTE, ADASYN, random undersampling, and near miss—highlighting the importance of these techniques in enhancing model accuracy and robustness.

Research conducted in [3] addresses the significant risk of bankruptcy among start-up companies through the utilization of advanced artificial intelligence techniques for predictive analysis. The study employs a range of algorithms, including CatBoost, RandomForest, XGBoost, AdaBoost, MLP, LogReg, and SVM, to categorize key features into critical credit aspects—such as Capacity, Capital, Collateral, Conditions, and Character—alongside the incorporation of Shapley values. This approach provides insights into factors predicting bankruptcy and enhances transparency with a ”glass-box” model. CatBoost achieved notable accuracy with a precision of 87%.

In [4], the study introduces a cutting-edge ensemble learning framework addressing corporate financial distress risks within the knowledge economy. Leveraging explainable artificial intelligence (XAI) techniques such as SHAP explanations and ICE interpretations, the research enhances transparency and accuracy in financial distress prediction. Empirical experiments using Polish company data reveal the GBoost model with random oversampling (ROS) as the top performer, offering insights into key indicators of distress through local PDP interpretations and actionable strategies for improvement via ICE interpretations. The study's global interpretation with SHAP's feature importance and interaction results aligns with financial experts' insights, emphasizing knowledge-driven innovation.

In [5], the study analyzes 5 banking companies' financial health with a modified Altman Zscore, examining key indicators like Working Capital to Total Assets (X1), Retained Earnings to Total Assets (X2), Earnings Before Interest Taxes to Total Assets (X3), and Market Value of Equity to

Book Value of Liabilities (X4). Findings indicate one company, Bank Mega tbk (MEGA), faces bankruptcy risk with an average Z-score of -3.32, while Bank Negara Indonesia (persero) tbk (BBNI), Bank Bukopin tbk (BBKP), Bank Mandiri (persero) tbk (BMRI), and Bank Rakyat Indonesia tbk (BBRI) demonstrate stability with Z-scores ranging from 4.38 to 4.62. This highlights the method's effectiveness in predicting corporate bankruptcy.

In [6], the research delves into the implementation of India's Insolvency and Bankruptcy Code 2016, aimed at modernizing the country's corporate insolvency regulations. Focusing on the Altman-Z score model, the study forecasts the safety levels of Indian banks regarding potential insolvency and bankruptcy. Results categorize banks into 'Safe,' 'Gray,' and 'Distress' zones based on Z-score thresholds, highlighting the robustness of Indian banks during the study period. Notable variations in bank performance were observed in the latter years, emphasizing the need for proactive management to mitigate insolvency risks.

In [7], the research assesses the performance of various machine learning models in predicting financial distress among listed companies in Vietnam. The study reveals that the extreme gradient boosting model achieved the highest accuracy of 95.66%, followed by the artificial neural network (ANN) with 91.68 accuracy. Through the application of SHAP values, the research identifies key financial indicators—such as long-term debts to equity, enterprise value to revenues, accounts payable to equity, and diluted EPS—that significantly influence distress predictions. The findings not only offer insights into the interpretability of black-box machine learning models but also provide credit rating companies with a new method to predict bond issuer default possibilities.

In [8], the research focuses on creating an ensemble model for predicting financial distress in companies within the Visegrad Group (V4) region. Utilizing real data from over 550,000 companies in Central Europe from the Amadeus database, the study trains and validates the model using 27 financial variables from 2016 to predict financial distress in 2017. The model identifies five significant predictors: current ratio, return on equity, return on assets, debt ratio, and net working capital. The hybrid model, combining RobustBoost, CART, and k-NN algorithms, achieves a superior accuracy of 94.25%, outperforming individual methods and offering a novel, high-performance tool for estimating financial distress risks in the V4 region.

In [9], the research focuses on developing a model to estimate the probability of corporate bankruptcy using various machine learning models. The study addresses the challenges of imbalanced data by employing data balancing techniques such as random undersampling and Synthetic Minority Over Sampling Technique (SMOTE). Using data from 2009 to 2013 on Poland manufacturing corporates and 64 selected financial indicators, the models—support vector machine (SVM), J48 decision tree, Logistic Model Tree (LMT), Random Forest (RF), and Decision Forest—are trained to predict bankruptcy occurrences. Results indicate significant improvement in predictive accuracy with machine learning techniques, particularly with SMOTE balancing. The Decision Forest model stands out with an impressive accuracy of 99%, followed by Random Forest at 98.7%, LMT at 93.8%, J48 decision tree at 92.3%, and SVM at 92%.

The work in[10] represents the analysis of financial risks associated with small and medium-sized businesses. The authors have utilized the Neural network method for the prediction and the proposed method is accurate in terms of prediction. The combined and compromise solution based approach for analysing financial risks for large businesses is presented in [11]. The proposed method follows the multiple attributes based decision and utilizes intuitionistic fuzzy sets. The paper [12], presents a neural network based method to analyse the financial fraud. Here, also neural network and Principal Component Analysis methods are employed. The paper presents the utilization of machine learning algorithms for financial analysis. Few more related work is presented in [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24],[25], [26]. The comparative analysis of some of the existing works is presented in Table 1.

The following are the research gaps identified after examining the recent relevant works.

- The some of the existing models lack clarity in explaining how financial indicators contribute to bankruptcy predictions, which limits its explainability.
- Due to the static nature, the majority of the models fail to adapt to changing economic conditions, which may result in decreased accuracy over time.
- An imbalanced dataset, with a bias towards the majority class (non-bankrupt instances), may lead to an underestimation of bankruptcy risks.
- The model's linear assumptions result in it assuming linear relationships between financial indicators and bankruptcy, potentially overlooking nonlinear patterns.
- By focusing solely on internal metrics, most of the models neglects external factors that could impact bankruptcy risk, contributing to its limited scope.

3 Proposed work

This study explores the critical domain of bankruptcy prediction in corporate finance, with the objective of enabling proactive financial management and strengthening business resilience amidst evolving challenges.

A visual representation of the sequential processes inherited in the proposed system is provided in Figure 1, offering a clear overview of its methodology and essential

Table 1: Comparative analysis of related work

Study	Methodology	Performance	Key Observations
[1]	Uses Kmeans-SMOTE method to mitigate data imbalance	SVM 99.1%, Naive Bayes 87%	enhance classification accuracy and mitigate data imbalance challenge
[2]	Applied five data balancing methods (random oversampling, SMOTE, ADASYN, Random undersam pling, and near miss)	Bagging 90.75%, SVML 80.06%, SVMR 79.96%, ANN 81.92%, RF 91.61%, Boosting 78.02%, KNN 59.72%, DT 81.43%, Logistic 65.60%	Represents significance of advanced machine learning techniques and proper data balancing
[3]	Top features are categorized into credit aspects (Capacity, Capital, Collateral, Conditions, Character), combined with Shapley values.	CatBoost (87%), RandomForest (83%), XGBoost (86%), AdaBoost (86%), MLP (65%), LogReg (59%), SVM (46%)	Design of the glass-box model enhances transparency
[4]	Local PDP and ICE methods were used for analysis	SHAP explanations for key indicators, ICE interpretations for improvement strategies	SHAP's results aligns with financial experts' insights
[5]	Z-score analysis for corporate bankruptcy prediction	Identified one unhealthy and four stable companies using a modified Altman Z-score	proposed model is more effective in predicting corporate bankruptcy
[6]	Altman-Z score model for Indian bank insolvency forecast	$Z > 2.99$ = "Safe," $1.81 < Z < 2.99$ = "Gray," $Z < 1.81$ = "Distress" zones	Emphasize on at modernizing the country's corporate insolvency regulations
[7]	Use various ML models for distress prediction	XGBoost (95.66%), Random Forest (85.35%), LogReg (86.23%), ANN (91.68%), DT (82.80%), SVM (87.89%)	Research findings offer insights into the interpretability of black-box machine learning models
[8]	Apply Ensemble methods for distress prediction in the regional context	RobustBoost (94.25%), CART (92.11%), k-NN (91.65%), Voting (92.69%), Avg Model (92.80%), Final Model (94.25%)	Ensemble approach presents a novel technique by offering an interpretability and high-performance
[9]	Apply ML models and also employ data balancing techniques	Decision Forest (99%), SVM (92%), J48 DT (92.3%), LMT (93.8%), RF (98.7%)	Significant improvement in predictive accuracy with ML techniques, and SMOTE balancing
[13]	Enhanced NN training for improved bankruptcy predictions	PSO (52.89%), MOA (72.24%), PSO-MOA (99.73%)	Hybrid approach enhances the speed of prediction and ensures higher accuracy
[14]	SAE technique for bankrupt firm prediction accuracy	SAE + Softmax Classifier achieves 87.9% accuracy for the Darden dataset and 98% accuracy for the Polish dataset	Better performance than traditional machine learning algorithms

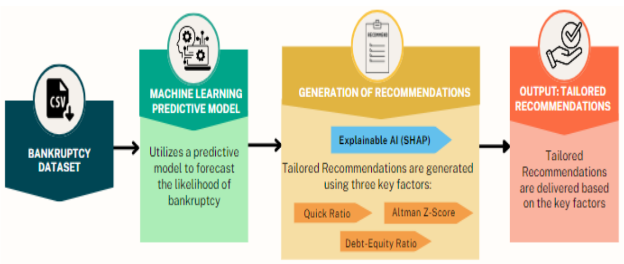

Figure 1: Architecture of proposed work

components. The architecture begins with the Machine Learning Predictive Model employs a robust framework for bankruptcy prediction by leveraging the US Company Bankruptcy Prediction Dataset. Initially, data imbalance is addressed through SMOTE oversampling, followed by data exploration, preprocessing, and normalization. This sets the stage for implementing six machine learning models, providing stakeholders with powerful tools for financial risk assessment. Based on the model's insights, tailored recommendations are generated to mitigate financial risks, focusing on the Altman Z-Score, Quick Ratio, and Debt-to-Equity Ratio, along with SHAP values for explainability. These recommendations guide companies in proactively managing bankruptcy risks and maintaining financial stability.

In this study, the main objective is to assess and compare the performance of multiple classifiers on imbalanced and balanced datasets using their default configurations. We aimed to establish a baseline performance for each model without introducing additional tuning complexity, thereby focusing our analysis on the impacts of class balancing via SMOTE rather than on optimized model settings. The US Company Bankruptcy Prediction Dataset [27], used in this study spans from 1999 to 2018 and includes financial data for 78,682 records of mid- and large-cap companies listed on the NYSE and NASDAQ. The dataset contains 18 financial indicators covering profitability, liquidity, and leverage, labeled as ”alive” or ”failed” to indicate solvency status. Companies come from diverse sectors, including technology, manufacturing, healthcare, finance, and consumer goods, making the dataset broadly representative and enhancing the model’s generalizability across industries and economic conditions.

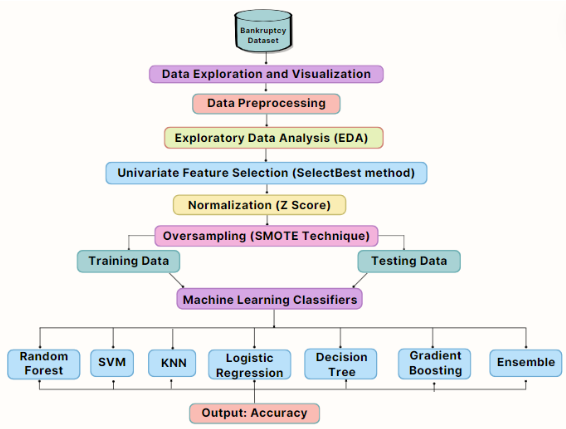

Figure 2: Workflow of bankruptcy prediction model

Illustrating the intricate process involved in building a dependable bankruptcy prediction model, Figure 2 emphasizes the essential steps critical for ensuring the model’s accuracy and reliability. It starts with data exploration and pre-processing phase that involves thorough examination and visual representation of the dataset. Essential preprocessing steps ensured data integrity by checking for missing values, summarizing numerical features, and encoding the target variable. Exploratory Data Analysis (EDA) followed, using a correlation matrix and heatmap to identify key feature relationships. Feature selection with SelectKBest highlighted the most relevant features, enhancing model performance. The dataset was then normalized with StandardScaler and balanced using SMOTE to address class imbalance. The resampled data was split into training (80%) and testing (20%) sets. Various classification algorithms, including Random Forest, Decision Tree, Logistic Regression, Gradient Boosting, SVM, KNN, and an Ensemble model, were trained and evaluated on these sets, with performance metrics calculated to compare their effectiveness.

Algorithm 1 outlines the steps required to clean and prepare the dataset, including handling missing values, encoding categorical variables, and normalizing the data o ensure the quality and integrity of the input data for model training. In the referred dataset, the only categorical feature is the target variable *status_label*, which has two possible values: ”alive” and ”failed.” We used binary encoding to represent these categories, mapping ”alive” to 1 and ”failed” to 0. Since this target variable is binary, it does not have high cardinality (a large number of unique categories). Therefore, high cardinality handling methods like target encoding, frequency encoding, or clustering categories are unnecessary for this dataset. Our simple binary encoding is both efficient and sufficient for this feature.

Algorithm 1 Data Preprocessing

Input: US Company Bankruptcy Prediction Dataset (78,682 rows, 19 attributes)
Output: Preprocessed Data
Start
Step 1: Load the Dataset D.
Step 2: Handle Missing Values

- For numerical values x_{ij} in feature j:
  - $x_{ij} = \begin{cases} \bar{x}_j & \text{if } x_{ij} \text{ is missing} \\ x_{ij} & \text{otherwise} \end{cases}$
- For categorical values x_{ij}:
  - $x_{ij} = \text{mode}(x_{\cdot j})$

Step 3: Encode Categorical Variables

- Convert categorical attributes to numerical values using one-hot encoding:
  - $x_{ij}(k) = \begin{cases} 1 & \text{if category } k \text{ matches} \\ 0 & \text{otherwise} \end{cases}$

Step 4: Adjust Column Names

- Rename columns for clarity and consistency.

Step 5: Conduct Exploratory Data Analysis (EDA)

- Visualize the distribution of the target variable.
- Summarize numerical features: compute mean $\bar{x}_j$, median $\tilde{x}_j$, and standard deviation σ_j for each numerical feature j.

Step 6: Encode Target Variable

- Convert 'status label' to numerical values:
  - $y_i = \begin{cases} 1 & \text{if 'alive'} \\ 0 & \text{if 'failed'} \end{cases}$

Step 7: Feature Selection

- Apply SelectKBest method with chi-square statistics: Select top k features based on chi-square test with target y.

Step 8: Normalize Data

- Scale features to zero mean and unit variance:
  - $x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$

End

Algorithm 2 outlines the method for addressing class imbalance in the dataset by generating synthetic samples for the minority class using techniques such as Synthetic Minority Over-sampling Technique (SMOTE). This approach helps to enhance the model’s performance by ensuring that the classifier does not become biased towards the majority class, thereby improving its ability to correctly predict bankruptcy cases. The algorithm begins by identifying the minority class samples, then creates synthetic samples by interpolating between existing minority samples. This process increases the representation of the minority class, leading to a more balanced dataset and, consequently, a more reliable predictive model. The proposed method selects the optimal number of nearest neighbors (k) by experimenting with various values and evaluating the impact on the performance of the model. It has been observed that k=5 provided the best balance between enhancing the minority class representation and maintaining model stability. Also, utilization of five neighbors yielded a more representative oversampling effect without introducing noise. Algorithm 3 covers the training of various

Algorithm 2 SMOTE Oversampling

Input: Preprocessed Data (with imbalanced classes)
Output: Balanced Data (after SMOTE oversampling)
Start
Step 1: Begin the SMOTE oversampling process
Step 2: Apply SMOTE (Synthetic Minority Over-sampling Technique)

- Set the sampling strategy parameter to 0.5.
- Generate synthetic samples for the minority class to balance the class distribution.

Step 3: Generate synthetic samples

- For each minority class example x_i:
  - Calculate the k nearest neighbors (using Euclidean distance or other distance metrics).
  - Randomly select one of the k nearest neighbors x_{zi}.
  - Generate a synthetic example x_{new} using the formula:

$$x_{\text{new}} = x_i + \lambda \times (x_{zi} - x_i)$$

where λ is a random number between 0 and 1.

Step 4: Repeat steps 1-3 until the desired balance between minority and majority classes is achieved
End

classification algorithms, including Random Forest, Decision Tree, Logistic Regression, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and an ensemble model. Their performance is evaluated using a range of metrics such as accuracy, precision, recall and F1 score. These metrics provide a comprehensive view of each model's ability to correctly predict bankruptcy cases while minimizing false positives and false negatives. The performance metrics of all models are then compared, and the best-performing model is selected based on its ability to balance accuracy, interpretability, and computational efficiency.

For the ensemble model, weighted voting approach is implemented that assigns weights based on individual model performance on testing data. This approach balanced model contributions, prioritizing models with higher accuracy. The models are selected based on their performance in preliminary testing and their complementary strengths in handling various data characteristics.

Algorithm 4 provides a comprehensive framework for analyzing financial metrics and generating tailored recommendations to mitigate bankruptcy risks. The process begins by utilizing the best-performing predictive model from previous algorithms to identify companies at risk of bankruptcy. This model leverages key financial indicators such as the Altman Z Score, Debt Equity Ratio, and Quick Ratio to assess the financial health of companies. In addition to generating recommendations, the algorithm employs SHAP (SHapley Additive exPlanations) analysis to provide model interpretability.

Algorithm 3 Classification Model Training

Input:
X: Feature matrix after preprocessing and SMOTE oversampling
y: Target variable vector after preprocessing and SMOTE oversampling
Output: Trained Models and Performance Metrics
Start
Step 1: Train Classification Algorithms
for each classification algorithm Alg_i (where $i = 1, 2, \ldots, N$) **do**
  Initialize Algorithm Alg_i
  Train Algorithm Alg_i on X_{train}, y_{train}
  Random Forest:
  - $y^{\text{RF}} = \text{RF}(X, \text{parameters})$

  Decision Tree:
  - $y^{\text{DT}} = \text{DT}(X, \text{parameters})$

  Logistic Regression:
  - $y^{\text{LogReg}} = \text{LogReg}(X, \text{parameters})$

  Gradient Boosting:
  - $y^{\text{GB}} = \text{GB}(X, \text{parameters})$

  Support Vector Machine (SVM):
  - $y^{\text{SVM}} = \text{SVM}(X, \text{parameters})$

  K-Nearest Neighbors (KNN):
  - $y^{\text{KNN}} = \text{KNN}(X, \text{parameters})$

  Ensemble Model (Simple Voting):
  - $y^{\text{Ensemble}} = \text{Ensemble}(X, \text{parameters})$

end for
Step 2: Predict on Testing Set
for each trained algorithm Alg_i **do**
  $y^{\text{test},i} = Alg_i(X_{\text{test}})$
end for
Step 3: Evaluate Performance Metrics

- Calculate performance metrics Metrics_i such as Accuracy, Precision, Recall, and F1-Score using $y^{\text{test},i}$ and y_{test}:
- **Accuracy:**
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
- **Precision:**
$$\text{Precision} = \frac{TP}{TP + FP}$$
- **Recall (Sensitivity):**
$$\text{Recall} = \frac{TP}{TP + FN}$$
- **F1-Score:**
$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Step 4: Select Best Model

- Compare the performance metrics Metrics_i for all algorithms and select the best-performing model.

End

4 Results and discussions

The major objective of the proposed method is to predict corporate bankruptcy by applying multiple machine learning classifiers on both imbalanced and balanced datasets. The tested algorithms included Random Forest, Decision Tree, Logistic Regression, Gradient Boosting, Support Vector Machine, K-Nearest Neighbors, and an Ensemble model. First, the model's performance is tested on the imbalanced dataset to understand the effects of class imbalance on accuracy, precision, recall, and F1-score. Then, the SMOTE was applied to address class imbalance by oversampling the minority class. The training process is repeated on this newly balanced dataset and comparative analysis of different classifiers, before and after applying SMOTE is performed. In the later phase, SHAP analysis id performed to generate tailored recommendations to mitigate corporate finance risks.

4.1 Experimental setup

The proposed method utilizes the bankruptcy distribution within the US Company Bankruptcy Prediction Dataset [27]. This dataset encompasses two target labels: "alive" for companies without bankruptcy history and "failed" for those that have undergone bankruptcy. Initially, the dataset comprised 73,462 instances labelled as "alive" and 5,220 instances labelled as "failed." To address dataset imbalance, SMOTE oversampling was employed, increasing "failed" instances to 36,731 and achieving a balanced dataset with 110,193 entries.

We have used an 80-20 train-test split of the dataset, with 80% of the data allocated for model training and 20% for testing. This split was applied uniformly across all machine learning models in our study. The focus is given on training and testing models using the single train-test split. This allowed us to analyze classifier performance on a stable partition of data while keeping the experimental process straightforward.

4.2 Experimental analysis

Illustrated in Figure 3 is the comparison of the target variable distribution before and after applying the SMOTE technique, highlighting the effective rebalancing achieved. This is crucial for ensuring robust performance and reliability in the bankruptcy prediction model. We did not observe signs of overfitting in our models. The performance metrics, such as accuracy, precision, recall, and F1-score, remained consistent across both training and test datasets, for each model. Given that our primary objective was a comparative analysis on imbalanced vs. balanced datasets, we relied on performance stability across the 80-20 train-test split to gauge generalization rather than implementing additional overfitting mitigation strategies.

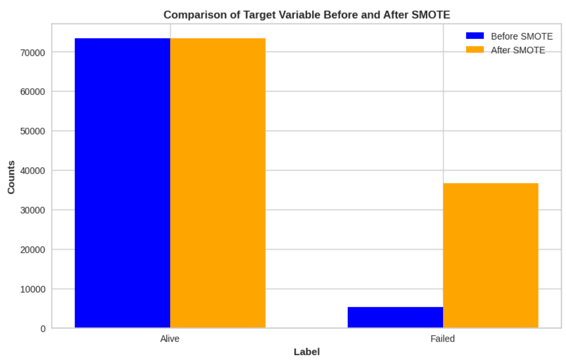

Figure 3: Before vs after SMOTE - target variable comparison

Table 2 summarizes the performance metrics (accuracy, precision, recall, F1-score) of various machine learning algorithms before applying SMOTE.

Table 2: Evaluation metrics of various models (before SMOTE)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	93.61	93.63	99.95	96.69
Decision Tree	89.74	94.66	94.31	94.48
Logistic Regression	93.21	93.23	99.97	96.48
Gradient Boosting	93.29	93.35	99.93	96.52
SVM	93.23	93.23	100.00	96.49
KNN	93.12	93.62	99.40	96.42
Ensemble	93.35	93.36	99.98	96.56

The performance evaluation graph in Figure 4 illustrates pre-SMOTE results, highlighting insights into machine learning classifier effectiveness on an imbalanced dataset. While all models show high accuracy and recall for the majority class ("alive"), precision varies notably. Models like Random Forest and Gradient Boosting exhibit superior precision compared to Logistic Regression and SVM, highlighting challenges in predicting the minority class ("failed"). Despite generally high F1-scores, indicating a balance between precision and recall, the findings underscore the need to tackle data imbalance.

Algorithm 4 Generation of Recommendations

Input:
Best Performing Model (from Algorithm 3)
Financial Metrics Analysis (including Altman Z-Score, Quick Ratio, Debt-to-Equity Ratio)
Output: Tailored Recommendations for Mitigating Bankruptcy Risks
Start

Step 1: Begin the process of generating recommendations

Step 2: Identify At-Risk Companies

- Utilize the best performing classification model (from Algorithm 3) to predict bankruptcy risk for companies.

Step 3: Analyze Financial Metrics

- **Altman Z-Score Calculation:**

$$Z = 1.2 \times X_1 + 1.4 \times X_2 + 3.3 \times X_3 + 0.6 \times X_4 + 1.0 \times X_5$$

where:

  - X_1: Working Capital / Total Assets
  - X_2: Retained Earnings / Total Assets
  - X_3: Earnings Before Interest and Taxes (EBIT) / Total Assets
  - X_4: Market Value of Equity / Book Value of Total Debt
  - X_5: Sales / Total Assets

- **Quick Ratio:**

$$\text{Quick Ratio} = \frac{\text{Current Assets} - \text{Inventory}}{\text{Current Liabilities}}$$

- **Debt-to-Equity Ratio:**

$$\text{Debt-to-Equity Ratio} = \frac{\text{Total Debt}}{\text{Total Equity}}$$

Step 4: Explain Model Decisions

- Employ Explainable AI techniques (e.g., SHAP values) to interpret model predictions and feature contributions.

Step 5: Generate Tailored Recommendations

- Based on the analysis:
  - Recommend strategies to improve financial health, such as reducing debt levels:

$$\text{Reduce Debt} = \text{Current Debt} - \text{Target Debt}$$

  - Suggest operational adjustments to enhance profitability and sustainability.

Step 6: Deliver Recommendations

- Present actionable insights and recommendations to stakeholders.

End

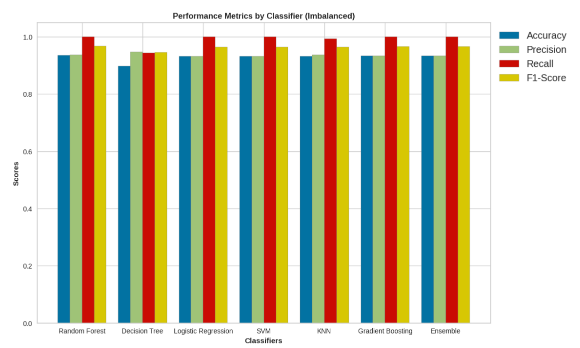

Figure 4: Evaluation metrics comparision (before SMOTE)

Table 3 displays the performance metrics of various models after applying SMOTE to balance the dataset.

Table 3: Evaluation metrics of various models (after SMOTE)

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	95.34	95.57	97.56	96.57
Decision Tree	86.74	91.52	88.37	89.92
Logistic Regression	67.41	67.48	98.99	80.25
Gradient Boosting	74.35	76.00	91.00	83.00
SVM	68.64	68.33	99.01	80.86
KNN	91.61	99.21	88.17	93.36
Ensemble	92.75	91.00	98.94	94.81

The performance evaluation graph in Figure 5 shows the results after applying the SMOTE technique and reveals several key insights into the effectiveness of various machine learning classifiers on a balanced dataset. Post-SMOTE, Random Forest emerges as the top performer across all metrics: accuracy, precision, recall, and F1-score. With an accuracy of 95.34%, precision of 95.57%, recall of 97.56%, and balanced F1-score of 96.57%, Random Forest proves its robustness and reliability. These results affirm its capability to accurately detect true positives (actual bankruptcies) while minimizing false positives, establishing it as the preferred model for predicting corporate bankruptcy.

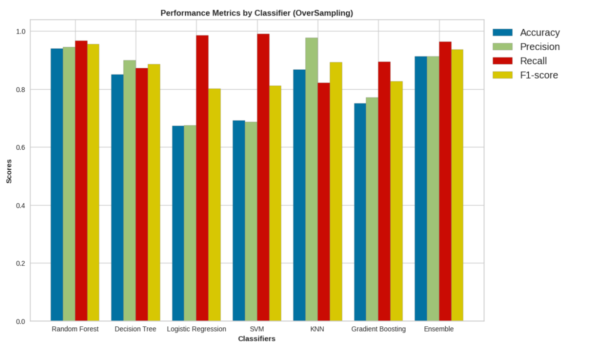

Figure 5: Evaluation metrics comparison (after SMOTE)

Utilizing Explainable AI (SHAP) as shown in Figure 6 enhanced model interpretability by detailing each feature's contribution to bankruptcy prediction, aiding stakeholders in comprehending factors influencing model decisions. The SHAP analysis identified key financial features like the Debt-to-Equity (D/E) Ratio, components of the Altman Z-Score (such as working capital and retained earnings), liquidity measures (e.g., current ratio), and profitability metrics (net income and gross profit) as consistently important predictors across models. These features align closely with established financial theories on bankruptcy, where high leverage, low liquidity, and declining profitability increase bankruptcy risk. By emphasizing these financially relevant features, the SHAP analysis bridges the model's predictions with traditional financial insights, enhancing interpretability and practical relevance for stakeholders.

Several key financial metrics in Table 4 were calculated to provide tailored recommendations for companies at risk of bankruptcy. The Quick Ratio assessed liquidity, the Debt-to-Equity Ratio evaluated financial leverage, and the Altman Z-Score gauged overall bankruptcy risk.

A thorough evaluation of SMOTE's impact on seven machine learning algorithms, is also performed. These analysis include, accuracy, precision, recall, and F1-score metrics across different test set proportions (ranging from 10% to 90%) and their corresponding metric values.

The analysis of the accuracy metric depicted in Figure 7 provides significant insight into how various machine learning algorithms perform when exposed to increasing proportions of testing data. Random Forest consistently exhibits high accuracy levels across all testing data proportions, emphasizing its stability and adaptability across diverse datasets.

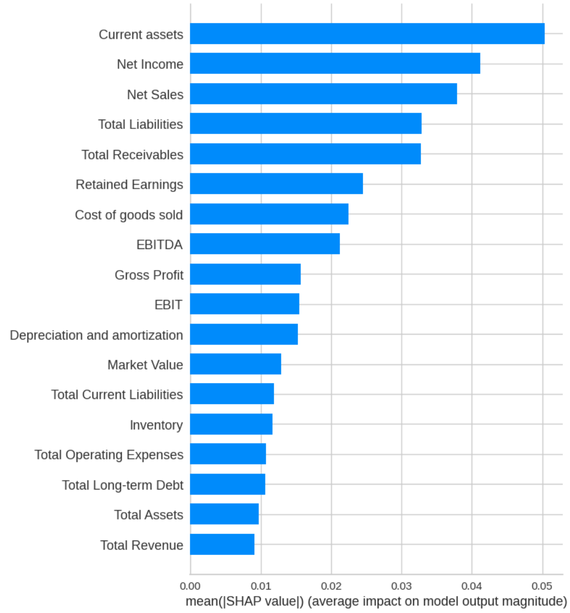

Figure 6: Explainable AI- SHAP analysis

Table 4: Recommendations for Financial Metrics

Metric	Recommendations
Quick Ratio	**Healthy -** High liquidity, low risk of short-term financial difficulties. **Acceptable -** Indicates moderate liquidity. **Cause for concern -** Low liquidity, potential difficulty in meeting short-term obligations.
Debt-to-Equity Ratio	**Conservative -** Indicates lower financial risk, less reliance on debt financing. **Moderate -** Balanced use of debt and equity financing. **Aggressive -** High financial risk, heavy reliance on debt financing.
Altman Z-Score	**Safe Zone –** Low Likelihood of Bankruptcy. **Grey Zone –** Moderate Risk of Bankruptcy. **Distress Zone –** High Likelihood of Bankruptcy.

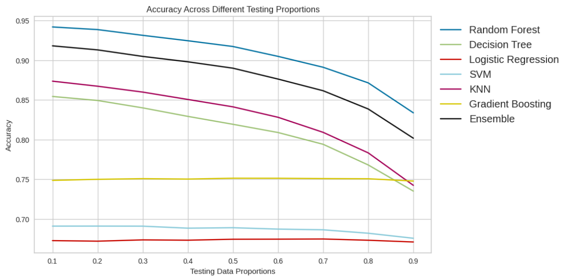

Figure 7: Accuracy of various classifiers

The analysis of the precision metric depicted in Figure 8 offers insights into machine learning algorithms' performance with increasing proportions of testing data. Initially, K-Nearest Neighbors (KNN) achieves the highest precision, followed by Random Forest and the Ensemble model. This trend underscores KNN's effectiveness in early identification of true positives. Random Forest and Ensemble models also maintain strong precision, highlighting their reliability across different testing data proportions.

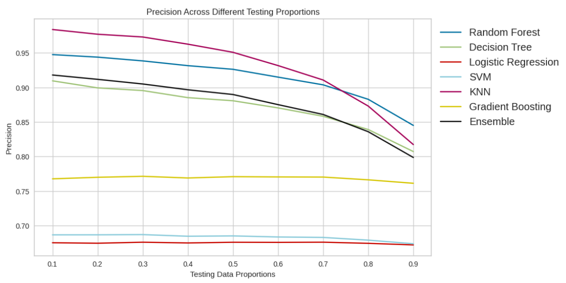

Figure 8: Precision of various classifiers

The analysis of the recall metric depicted in Figure 9 provides insights into machine learning algorithms' performance with increasing testing data proportions. Initially, Support Vector Machine (SVM) and Logistic Regression show notable increases in recall, indicating their effectiveness in identifying true positive instances. As testing data proportions increase, Ensemble and Random Forest models also demonstrate commendable rises in recall levels, particularly in the early testing phases, highlighting their reliability across diverse data splits. The Figure 10 illustrates how F1-scores evolve across different testing proportions, providing insights into each algorithm's performance under varying data splits. Random Forest and Ensemble models show significant increases in F1-scores initially, indicating their ability to balance precision and recall effectively. KNN and Decision Tree also exhibit notable rises in F1-scores during early testing phases, reflecting their capability to maintain balanced performance as the dataset size increases.

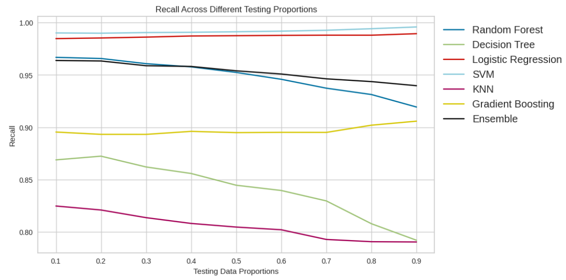

Figure 9: Recall of various classifiers

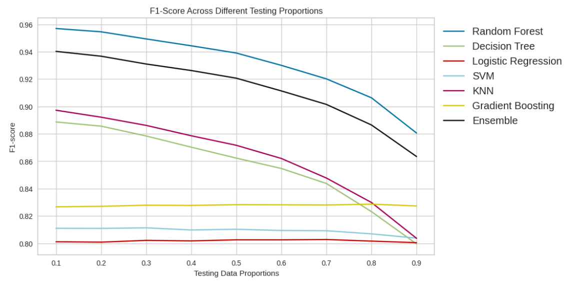

Figure 10: F1-Score of various classifiers

4.3 Performance analysis

This section elaborates the comparative analysis of proposed work and some of the existing works. Table 5 compares various research works and their incorporation of key methodologies. It highlights the comprehensive approach of the proposed work, which uniquely integrates all critical methodologies for effective bankruptcy prediction and provide tailored recommendations.

Table 5: Comparison of proposed work and existing research works

Author	Classification Model	SHAP	Altman Z score	SMOTE
[2]	✓	✗	✗	✓
[4]	✗	✓	✗	✗
[6]	✗	✗	✓	✗
[7]	✓	✓	✗	✗
[Proposed]	✓	✓	✓	✓

Accuracy measures the overall correctness of the model by calculating the proportion of true results (both true positives and true negatives) among the total number of cases examined. As shown in Figure 11, XGBoost [7] has accuracy of 95.66% , indicating that it correctly classifies most instances. The proposed method uses RF model and demonstrates high accuracy of 95.34%, reflecting its robustness and reliability.

Precision (Positive Predictive Value, PPV) measures the proportion of positive identifications that were actually correct. As given in Figure 12 , ANN [2] has a precision of 99.79% suggesting it accurately identifies actual positives with minimal false positives. The proposed method with KNN has a precision of 99.21%, indicating its effectiveness in maintaining low false positive rates.

Recall (Sensitivity, True Positive Rate) measures the pro-

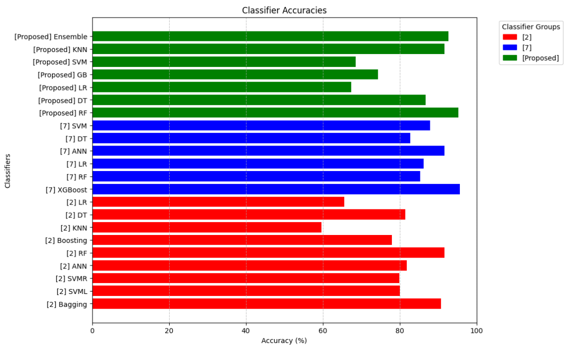

Figure 11: Comparative analysis with respect to accuracy

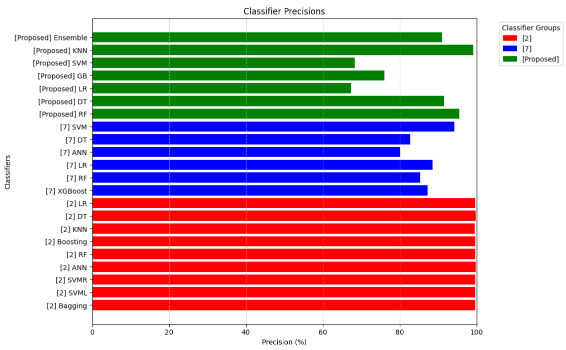

Figure 12: Comparative analysis with respect to precision

portion of actual positives that were correctly identified by the model. From Figure 13, it can be observed that proposed method (SVM model) with a recall of 99.01% excels in identifying almost all actual positive cases, which is crucial for detecting bankruptcy risks. Also, the proposed method with RF model and Ensemble model shows a high recall of 97.56% and 98.94% respectively, ensuring that most true positives are identified.

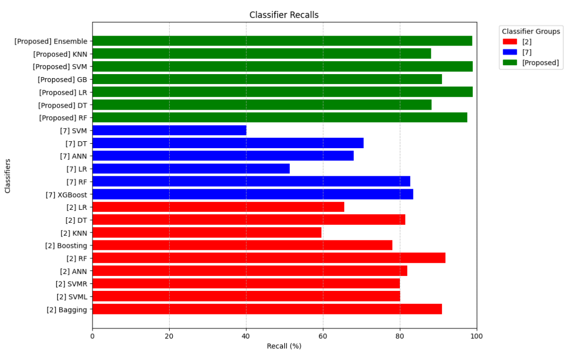

Figure 13: Comparative analysis with respect to recall

The F1 Score is the harmonic mean of precision and recall, providing a balance between the two metrics. As shown in Figure 14, the proposed method with RF model has F1 score of 96.57% demonstrates exceptional balance and overall performance, making it a strong choice for predicting corporate bankruptcy. Also the proposed Ensemble model has F1 score of 94.81% that shows a well-rounded performance.

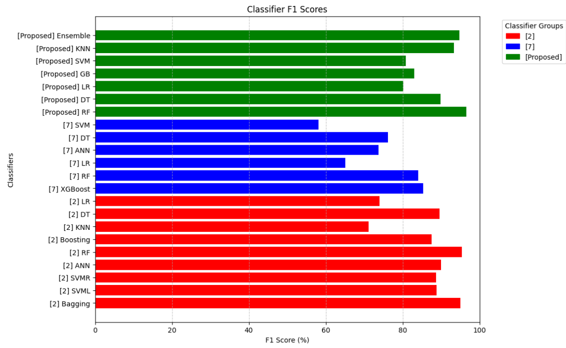

Figure 14: Comparative analysis with respect to F1-score

A primary constraint associate with the proposed methodology is its reliance on historical financial indicators, which may not fully account for sudden economic shifts or unprecedented events like the COVID-19 pandemic. This dependence limits the model's predictive accuracy in capturing abrupt financial distress signals, as it assumes historical patterns will continue in the future. Future work could explore integrating external economic indicators or real-time data sources to address these limitations and improve adaptability to sudden market shifts.

5 Conclusion

In conclusion, the proposed approach represents a significant advancement in the field of bankruptcy prediction, effectively addressing critical limitations observed in existing models. Through the strategic integration of advanced machine learning techniques, our methodology introduces a flexible and responsive system capable of accurately assessing bankruptcy risk. Key financial metrics such as the Altman Z-Score, Quick Ratio and Debt-to-Equity Ratio, complemented by Explainable AI Techniques, enhance transparency and interpretability, providing stakeholders with valuable insights for informed decision-making regarding corporate financial health.

Looking ahead, there is ample room for further refinement and optimization of the machine learning model to achieve even greater predictive accuracy. This entails exploring additional financial indicators and economic factors to strengthen the robustness of the bankruptcy prediction system. Additionally, alternative machine learning algorithms and ensemble methods will be thoroughly investigated to compare and optimize performance, ensuring the continued effectiveness of the proposed approach in real-world scenarios. Such endeavours are crucial for maintaining the relevance and practical applicability of the methodology in addressing the evolving challenges of financial risk management across diverse industries.

References

[1] Maulana, Didit Johar, Siti Saadah, and Prasti Eko Yunanto. Kmeans-SMOTE Integration for Handling Imbalance Data in Classifying Financial Distress Companies using SVM and Naïve Bayes. Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi) 8(1): 54-61,2024. https://doi.org/10.29207/resti.v8i1.5140

[2] Liashenko, Olena, Tetyana Kravets, and Yevhenii Kostovetskyi. Machine learning and data balancing methods for bankruptcy prediction. Ekonomika 102(2): 28-46, 2023. https://doi.org/10.15388/Ekon.2023.102.2.2

[3] Enkhtuya, Tuguldur, and Dae-Ki Kang. Bankruptcy Prediction with Explainable Artificial Intelligence for Early-Stage Business Models. International Journal of Internet, Broadcasting and Communication 15(3):58-65, 2023.

[4] Fan, Mengting, Zan Mo, Qizhi Zhao, and Zhouyang Liang. Innovative Insights into Knowledge-Driven Financial Distress Prediction: a Comprehensive XAI Approach. Journal of the Knowledge Economy, 15: 12554–12595, 2024. https://doi.org/10.1007/s13132-023-01602-4

[5] Wahyuni, Sari. Prediction of Bankruptcy Levels Using the Almant Z-Method Score in Banking Companies on the Indonesia Stock Exchange for the 2018-2021 Period. MANKEU (Jurnal Manajemen Keuangan), 1(2), 2023. https://doi.org/10.61167/mnk.v1i2.39.

[6] Lokeshnath, B., and M. Sandhya. Insolvency And Bankruptcy Code: A Study of Indian Banks With Reference To Altman Z Score. EPRA International Journal of Economic and Business Review (JEBR),11(8):69-79, 2023. https://doi.org/10.36713/epra14205

[7] Tran, Kim Long, Hoang Anh Le, Thanh Hien Nguyen, and Duc Trung Nguyen. Explainable machine learning for financial distress prediction: evidence from Vietnam. Data, 7(11):160, 2022.

[8] Pavlicko, Michal, Marek Durica, and Jaroslav Mazanec. Ensemble model of the financial distress prediction in Visegrad group countries. Mathematics, 9(16):1886,2021.

[9] Alam, Talha Mahboob, Kamran Shaukat, Mubbashar Mushtaq, Yasir Ali, Matloob Khushi, Suhuai Luo, and Abdul Wahab. Corporate bankruptcy prediction: An approach towards better corporate world. The Computer Journal, 64(11):1731-1746, 2021. https://doi.org/10.1093/comjnl/bxaa056

[10] Xiaohui Wang, Research on Financial Risk Predictio[n and Prevention for Small and Medium-Sized Enterprises - Based on a Neural Network. Informatica, 47:153–160, 2023. https://doi.org/10.31449/inf.v47i8.4884

[11] Huanwen Liu, Enhanced CoCoSo Method for Intuitionistic Fuzzy MAGDM and Application to Financial Risk Evaluation of High-Tech Enterprises. Informatica, 48:1–14, 2024. https://doi.org/10.31449/inf.v48i5.5169

[12] Xiaohui Wang. Research on Financial Risk Prediction and Prevention for Small and Medium-Sized Enterprises - Based on a Neural Network. Informatica, 47:153–160, 2023. https://doi.org/10.31449/inf.v47i8.4884

[13] Zilu Liang, Yunji Liang. A Study of Identification of Corporate Financial Fraud Using Neural Network Algorithms in an Information-Based Environment. Informatica, 47:165–172, 2023. https://doi.org/10.31449/inf.v47i9.5220

[14] Soui, Makram, Salima Smiti, Mohamed Wiem Mkaouer, and Ridha Ejbali. Bankruptcy prediction using stacked auto-encoders. Applied Artificial Intelligence,34(1):80-100, 2020. https://doi.org/10.1080/08839514.2019.1691849

[15] Fares, Omar H., Irfan Butt, and Seung Hwan Mark Lee. Utilization of artificial intelligence in the banking sector: a systematic literature review. Journal of Financial Services Marketing, 28(4):835-852, 2023. https://doi.org/10.1057/s41264-022-00176-7

[16] Hanif, Ambreen. Towards explainable artificial intelligence in banking and financial services. arXiv preprint arXiv:2112.08441, 2021.

[17] M, Hiran, G, Megavarshini, Sreenivasan, Aswathy, Suresh, Ma. Machine Learning in the Banking Sector, 2023. 1081-1088. 10.46254/AN13.20230313.

[18] Bahoo, Salman, Marco Cucculelli, Xhoana Goga, and Jasmine Mondolo. Artificial intelligence in Finance: a comprehensive review through bibliometric and content analysis. SN Business & Economics, 4(2): 23, 2024. https://doi.org/10.1007/s43546-023-00618-

[19] Sharma, Bhumiswor, P. Srikanth, and S. Jeevananda. Financial Distress and Value Premium using Altman Revised Z-score Model. Vision, 2023. https://doi.org/10.1177/09722629231198604

[20] Cındık, Zeynep, Armutlulu, İsmail. A revision of Altman Z-Score model and a comparative analysis of Turkish companies' financial distress prediction. National Accounting Review, 3.:237-255, 2021. 10.3934/NAR.2021012

[21] Chen, Tsung-Kang, Hsien-Hsing Liao, Geng-Dao Chen, Wei-Han Kang, and Yu-Chun Lin. Bankruptcy

prediction using machine learning models with the text-based communicative value of annual reports. Expert Systems with Applications 233: 120714, 2023. https://doi.org/10.1016/j.eswa.2023.120714

[22] Kim, Hyeongjun, Hoon Cho, and Doojin Ryu. Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data. Computational Economics, 59(3):1231-1249, 2022. https://doi.org/10.1007/s10614-021-10126-5

[23] Dablain, Damien, Krawczyk, Bartosz, Chawla, Nitesh. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. IEEE Transactions on Neural Networks and Learning Systems:1-15, 2022. 10.1109/TNNLS.2021.3136503

[24] Elreedy, Dina, Amir F. Atiya, and Firuz Kamalov. A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning. Machine Learning, 113: 4903–4923, 2024. https://doi.org/10.1007/s10994-022-06296-4

[25] Hairani, Hairani, Khurniawan Eko Saputro, and Sofiansyah Fadli. K-means-SMOTE untuk menangani ketidakseimbangan kelas dalam klasifikasi penyakit diabetes dengan C4. 5, SVM, dan naive Bayes. Jurnal Teknologi dan Sistem Komputer, 8(2):89-93, 2020. https://doi.org/10.14710/jtsiskom.8.2.2020.89-93

[26] Salehi, Amir Reza, and Majid Khedmati. A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data. Scientific Reports, 14(1):5152, 2024. https://doi.org/10.1038/s41598-024-55598-1

[27] Kaggle, US Company Bankruptcy Prediction Dataset. https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset