Optimizing Bankruptcy Prediction on Imbalanced Data using XGBoost with Random Oversampling and Chi-Square

Revalina Suyatno [1]*, Erika Devi Udayanti [2]*, Ika Novita Dewi [3]*
* Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang
112202206897@mhs.dinus.ac.id [1], erikadevi@dsn.dinus.ac.id [2], ikadewi@dsn.dinus.ac.id [3]

Article Info

Article history:

Received 2025-11-24
Revised 2026-01-06
Accepted 2026-01-08

Keyword:

Bankruptcy Prediction,
Data Imbalance,
Random Oversampling (ROS),
Chi-Square,
Extreme Gradient Boosting (XGBoost).

ABSTRACT

In the midst of modern financial dynamics, the ability to predict corporate bankruptcy holds strategic significance, as it directly affects economic stability and investor confidence. However, the development of a reliable predictive model is often hindered by the complex nature of financial data, particularly the class imbalance between bankrupt and non-bankrupt companies. This imbalance causes models to become biased toward the majority class, thereby reducing their sensitivity in detecting bankruptcy cases which are, in fact, the most critical for financial decision-making. This research aims to construct a more balanced and sensitive bankruptcy prediction model by specifically addressing the issue of data imbalance. The proposed approach integrates the Random Oversampling (ROS) technique to equalize class distribution, Chi-Square feature selection to identify the most informative financial variables, and the Extreme Gradient Boosting (XGBoost) algorithm as the core predictive model. The dataset used is the UCI Taiwanese Bankruptcy Prediction dataset, consisting of 6,819 observations and 96 financial ratio variables. Experimental results show that the Chi-Square method successfully identified 20 influential variables, including Per Share Net Profit Before, Debt Ratio, and ROA(B) Before Interest and Depreciation After Tax. The proposed XGBoost model achieved an overall accuracy of 0.9648 and an F1-score of 0.4286, demonstrating superior performance. These findings confirm that the combination of ROS, Chi-Square, and XGBoost effectively enhances data balance and prediction sensitivity for the bankruptcy class. This research is expected to serve as a foundation for developing financial decision-support systems capable of providing early warnings of potential corporate bankruptcy.

I. Introduction

Corporate bankruptcy prediction has become a strategic issue in modern financial management, as it directly affects national economic stability, business sustainability, and investor confidence [1], [2]. Each year, thousands of companies across various sectors face bankruptcy, resulting in massive job losses and significant economic downturns. Amid the global market dynamics characterized by digital disruption, currency fluctuations, and regulatory changes many firms experience liquidity pressures that escalate into increasingly complex financial distress risks [3], [4]. Historically, a company's financial condition has often been assessed through multiple ratios derived from annual financial statements [5]. This approach is considered practical because it provides a general overview of corporate performance in terms of profitability, liquidity, and solvency. However, as business environments become more complex and analytical precision grows in importance, researchers have begun to question the effectiveness of these traditional approaches in identifying potential bankruptcy risks [6].

To overcome the limitations of conventional approaches, various bankruptcy prediction methods have been developed in recent decades. Classical statistical models such as Altman Z-Score, Springate, and Grover G-Score marked the early foundations of quantitative assessment for bankruptcy risk

[7], [8]. However, these traditional approaches are limited in their ability to capture complex non-linear interactions among financial variables [9]. With the advancement of computational technology, machine learning techniques have increasingly been adopted for bankruptcy prediction. Algorithms such as Support Vector Machine (SVM), Random Forest, and Multilayer Perceptron (MLP) have demonstrated improved predictive accuracy and generalization capability [10], [11], [12]. Nevertheless, most of these models still face substantial challenges when dealing with imbalanced datasets, where the proportion of non-bankrupt companies significantly exceeds that of bankrupt ones. This imbalance often leads to bias toward the majority class, thereby reducing the model's ability to effectively identify firms with a high risk of failure [13].

The issue of data imbalance remains one of the major challenges in financial classification modelling [14], [15]. When the proportion of bankrupt firms is significantly smaller than that of non-bankrupt ones, the learning process tends to be biased toward the majority class. To address this issue, an effective data balancing strategy is essential to enhance the model's sensitivity toward minority instances [16]. One commonly applied approach is the Synthetic Minority Oversampling Technique (SMOTE), which generates artificial samples to increase the representation of the minority class [17]. However, synthetic generation may sometimes produce instances that do not accurately reflect real financial conditions, potentially distorting the intrinsic data distribution. As an alternative, Random Oversampling (ROS) provides a simpler yet more reliable solution by duplicating minority samples while maintaining the authenticity of the original dataset [18], [19]. This method allows the model to learn from a more balanced class distribution without introducing the potential noise often associated with synthetic data.

In addition to class balancing, another crucial aspect in building a reliable bankruptcy prediction model is feature selection [20]. The primary objective of this process is to identify the most relevant variables that can enhance both the efficiency and predictive accuracy of the model [21]. In this research, the Chi-Square method is employed to evaluate the statistical dependence between each independent variable and the target label. This approach is particularly suitable for financial datasets that consist of a mix of numerical and categorical features, as it effectively quantifies the significance level of each attribute. By filtering out only the most informative features, the Chi-Square method helps reduce model complexity, lower computational overhead, and accelerate the training process without compromising critical information [22].

After the dataset was optimized through class balancing and feature selection, the next crucial stage involved constructing the predictive model. This research adopts the Extreme Gradient Boosting (XGBoost) algorithm as the core classifier due to its proven capability in handling large-scale and non-linear financial data [23]. XGBoost is an ensemble learning technique that iteratively combines multiple weak learners to form a robust predictive model. By leveraging the principle of gradient boosting, the algorithm continuously refines prediction errors from previous iterations while incorporating a regularization mechanism to prevent overfitting, [24]. Compared to other algorithms such as Random Forest and Logistic Regression, XGBoost demonstrates superior computational efficiency and more consistent predictive accuracy, particularly when applied to financial datasets characterized by complex inter-variable relationships [25], [26].

Several prior researches have attempted various approaches to model corporate bankruptcy; however, each still presents notable limitations. The research by Kaya et al. (2022) [27] employing a MLP model achieved an accuracy of approximately 80% in identifying corporate financial conditions. Although this neural network–based approach effectively captured complex financial patterns, it failed to adequately address class imbalance, which reduced its robustness in real-world scenarios. Rahayu et al. (2023) [15] emphasized that data imbalance remains a key factor contributing to reduced predictive accuracy in machine learning models, even when algorithms such as Logistic Regression and SVM are used. Furthermore, Kristanti et al. (2023) [28] demonstrated that ensemble methods, including Random Forest, can improve predictive stability, but their performance declines significantly when the minority class is underrepresented and no balancing technique is applied. Maulana et al. (2024) [14] proposed a hybrid approach to handle imbalance issues; however, their findings indicated that applying SMOTE-based synthetic sampling tends to increase overfitting and limit the model's generalization ability. Based on these researches, there remains a critical need for an approach that systematically integrates class balancing, statistical feature selection, and a comprehensive learning framework to enhance predictive accuracy and model reliability.

Although ROS, Chi-Square, and XGBoost have been individually applied in prior prediction researches, this research contributes by systematically analyzing their combined and sequential impact within a unified experimental framework. Instead of proposing a new algorithm, this research focuses on evaluating how each methodological stage class balancing and feature selection affects predictive performance when applied to the same dataset. By conducting stage-wise comparisons using the Taiwanese Company Bankruptcy Dataset, this research provides empirical insights into the trade-offs between accuracy and sensitivity in imbalanced financial data. This analytical approach distinguishes the research from previous works that often assess these techniques in isolation.

II. METHOD

This research is designed to develop an effective and reliable corporate bankruptcy prediction model through the integration of oversampling techniques and feature selection within an ensemble-based machine learning framework. In general, the research stages include: (1) data preparation and understanding, (2) Exploratory Data Analysis (EDA), (3) data preprocessing, which involves normalization and handling of class imbalance, (4) feature selection using the Chi-Square method, (5) predictive model development using the XGBoost algorithm, and (6) model performance evaluation. The complete research methodology is illustrated in Figure 1, which presents a systematic process starting from data preparation, class balancing, feature selection, and XGBoost modeling, to the final model performance evaluation.

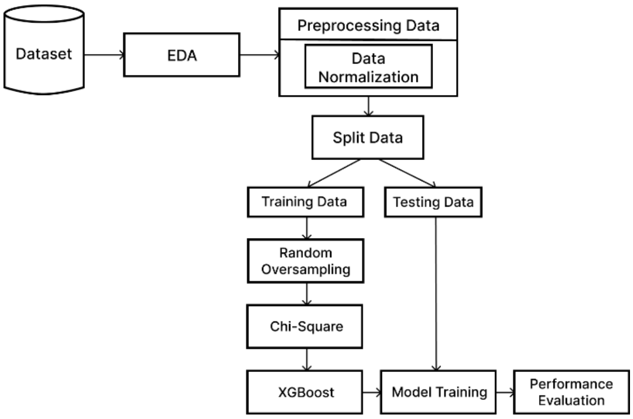

Figure 1. Research Flow

A. Dataset

This research employs the Taiwanese Company Bankruptcy Dataset [29], which is a publicly available dataset designed to predict the potential bankruptcy of companies based on various financial indicators. The dataset contains a total of 6,819 corporate entities from different fiscal years and includes 96 numerical attributes, along with a single target variable indicating the company's bankruptcy status.

Each attribute represents a financial ratio that reflects the company's financial health and operational performance. The key categories of these attributes include: (1) Profitability indicators, such as Return on Assets (ROA) before and after tax, Operating Profit Rate, and Net Income to Total Assets, which assess a firm's efficiency in generating profit. (2) Liquidity measures, such as Current Liability to Current Assets, which evaluate the company's ability to meet short-term obligations. (3) Leverage metrics, including Liability to Equity and Equity to Liability, which describe the firm's capital structure and financial risk. (4) Cash flow indicators, such as Cash Flow to Total Assets and Cash Flow from Operations (CFO) to Assets, which illustrate the relationship between cash generation and asset utilization. (5) Flag variables, such as Liability, Assets Flag and Net Income Flag, which signal extreme conditions or potential financial anomalies.

In addition, the dataset includes dozens of other financial ratios that collectively provide a comprehensive representation of a company's financial health and capital structure. The target variable in this dataset is labeled "Bankrupt?", where a value of 1 indicates that the company is bankrupt, and 0 denotes that the company is financially stable (non-bankrupt). An overview of the dataset's structure is presented in Table I.

TABLE I
DATASET OVERVIEW

Bankrupt?	**ROA(C) before interest and depreciation before interest**	**ROA(A) before interest and % after tax**	…	**Equity to Liability**
1	0.370594	0.424389	…	0.016469
1	0.464291	0.538214	…	0.020794
1	0.426071	0.499019	…	0.016474
0	0.390922	0.445704	…	0.015663
0	0.508361	0.570922	…	0.034888
0	0.488519	0.545137	...	0.065826

The "Bankrupt?" column serves as the target variable. The subsequent columns contain key financial ratios that act as predictor variables, such as Return on Assets (ROA), Equity to Liability, and other ratios that collectively describe the company's overall financial condition.

B. Exploratory Data Analysis (EDA)

After the data preparation stage, an EDA was conducted to understand the characteristics, patterns, and potential anomalies within the dataset. This phase involved performing descriptive statistical analysis, assessing data completeness, and exploring intervariable relationships to identify underlying financial patterns that could influence the prediction of bankruptcy.

The first step involved conducting a data completeness analysis to ensure that no missing values were present, as such omissions could affect the reliability of the model training process. Subsequently, the distribution of the target variable "Bankrupt?" was analyzed using a bar chart generated through the Seaborn library. This visualization served to identify the degree of class imbalance between bankrupt and non-bankrupt companies forming the analytical foundation for selecting an appropriate data balancing strategy in the following stage.

The next stage involved a feature correlation analysis using the Pearson correlation coefficient (r) to measure the strength of linear relationships among numerical variables. The results of this analysis were then visualized in the form of a correlation heatmap, which illustrates the degree of association between features and helps identify potential multicollinearity within the dataset.

C. Data Normalization

In this research, a data normalization process was applied to ensure that all features operate on a comparable scale. This step is essential because variations in the numerical range among variables can cause certain features to dominate others, thereby affecting the stability of the model's learning process. One of the most widely adopted methods is Min-Max Scaling, which transforms feature values into a standardized range between 0 and 1. This approach has proven effective for tree-based algorithms such as XGBoost, as it enhances convergence speed and maintains numerical stability during model training. The normalization process was computed using Equation (1).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad (1)$$

Where:
x' = the normalized value,
x = the original feature value,
x_{min} = the minimum value of the feature, and
x_{max} = the maximum value of the feature.

D. Split Data

After the normalization process, the dataset was divided into training and testing subsets using a stratified hold-out validation strategy. The training data were used to build and optimize the model parameters, while the testing data served to evaluate the model's generalization capability on previously unseen data. The dataset was split into 80% training data and 20% testing data to preserve the original class distribution. To prevent data leakage, the scaling process was fitted exclusively on the training data, while the testing data were only transformed using the parameters learned from the training set.

Mathematically, the dataset partitioning process can be expressed in Equation (2).

$$D = D_{train} \cup D_{test} , D_{train} \cup D_{test} = \emptyset \qquad (2)$$

This stage ensures that the training and evaluation processes are conducted independently, allowing the resulting performance metrics to accurately reflect the true predictive capability of the model.

E. Random Oversampling (ROS) for Handling Data Imbalance

The dataset used in this research exhibits a significant class imbalance, with bankrupt companies representing the minority class. To mitigate model bias and enhance sensitivity toward the minority class, ROS technique was applied. ROS operates by randomly duplicating samples from the minority class until a balanced proportion with the majority class is achieved, ensuring that both classes contribute equally during model training.

Although ROS may increase the risk of overfitting due to the duplication of minority-class samples, several mitigation strategies were incorporated into the modeling pipeline. First, oversampling was applied exclusively to the training set after data splitting, ensuring that the test set remained untouched and representative of real-world class distributions. Second, XGBoost regularization mechanisms, including L2 regularization (reg_lambda), controlled tree depth (max_depth), and subsampling of both observations and features (subsample and colsample_bytree), were employed to limit model complexity. These constraints reduce the model's tendency to memorize duplicated samples and encourage more generalized decision boundaries. In addition, this research also evaluated SMOTE as a comparative approach. However, ROS was selected as the primary resampling strategy due to its ability to preserve the original distribution of financial features, thereby avoiding the introduction of synthetic patterns that may distort the underlying data structure.

The conceptual formulation of the oversampling process is represented in Equation (3).

$$D' = D_{majority} \cup \{x_i | x_i \in D_{minority}\}^k \qquad (3)$$

Where:
D' = the oversampled dataset,
$D_{majority}$ = the set of samples from the majority class,
$D_{minority}$ = the set of samples from the minority class, and
k = the replication factor, representing the number of times minority class samples are duplicated until a balanced class distribution is achieved.

Through this approach, the class distribution becomes more balanced, enabling the model to learn patterns from both classes proportionally without bias toward the dominant class.

F. Chi-Square Feature Selection

To reduce data complexity and retain only the most informative features, this research employs the Chi-Square method as a feature selection technique. In the context of bankruptcy prediction, the primary objective of feature selection is to identify variables that provide strong and meaningful signals regarding the likelihood of bankruptcy. This enables the model to focus on the most relevant attributes while minimizing the influence of redundant or noisy features that could otherwise degrade prediction accuracy and interpretability.

The Chi-Square method quantifies the statistical dependence between each feature and the binary target label "Bankrupt?". In essence, it evaluates how far the observed frequency (O_i) deviates from the expected frequency (E_i) under the assumption that the feature is independent of the target variable. The Chi-Square test value (x^2) for a given feature can be formulated as shown in Equation (4).

$$x^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \qquad (4)$$

Features with higher x^2 values indicate a stronger contribution to class differentiation and are therefore prioritized for retention in the modelling process. Through this approach, the Chi-Square method effectively eliminates less relevant attributes, reduces feature dimensionality, accelerates the model training process, and often enhances the model's generalization capability.

Since the financial ratios in the dataset are numerical in nature, a normalization process using Min–Max scaling is applied prior to Chi-Square computation to ensure that all feature values are non-negative, which is a key requirement for the validity of the Chi-Square test. This preprocessing step allows numerical financial indicators to be appropriately evaluated within the Chi-Square framework without altering their relative distributions.

Furthermore, the number of selected features is limited to the top 20 variables with the highest Chi-Square scores. This threshold is chosen to achieve a balance between preserving sufficient financial information and reducing model dimensionality, thereby improving computational efficiency and mitigating the risk of overfitting. Selecting a fixed number of top-ranked features also facilitates a more stable and interpretable model structure, particularly when combined with a high-capacity classifier such as XGBoost.

G. Extreme Gradient Boosting (XGBoost) Modelling

XGBoost is an Ensemble Learning algorithm based on Decision Trees that constructs models in a gradual manner through iterative boosting. Each newly generated tree aims to correct the prediction errors made by the previous trees, resulting in a final model that represents a calibrated combination of multiple weak learners, collectively producing strong and stable predictions. Mathematically, the objective function of XGBoost is formulated in Equation (5).

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \mho(f_t) \tag{5}$$

where $l(y_i, \hat{y}_i)$ represents the loss function that measures the deviation between the predicted and actual values, while $\mho(f_t)$ denotes the regularization term that controls the complexity of each tree at iteration t.

The configuration of XGBoost parameters plays a crucial role in optimizing the model's performance and robustness. Table II summarizes the key parameters determined through grid search along with the rationale behind each choice. The number of estimators (n_estimators) is set to 300 to balance predictive accuracy and computational efficiency, ensuring sufficient model learning without excessive training time. The maximum tree depth (max_depth) is configured to 7 to control model complexity, prevent overfitting on the training data, and enhance generalization capability. A learning_rate of 0.2 is selected to allow gradual and stable weight updates, thereby minimizing the risk of overshooting the optimal solution.

To further enhance robustness and mitigate overfitting, the *subsample* and colsample_bytree parameters are each set to 0.7, meaning that every tree is trained using a random subset of the available data samples and features. Finally, L2 regularization (reg_lambda) is applied with a value of 1 to reduce model variance and improve generalization. This combination of parameters has been empirically found to deliver a balanced and stable performance on the bankruptcy dataset utilized in this research.

TABLE II
OPTIMAL XGBOOST PARAMETER CONFIGURATION

Parameters	Value	Description
n_estimators	300	Balancing model accuracy and computational efficiency
max_depth	7	Controlling model complexity to prevent overfitting.
learning_rate	0.2	Gradual and stable weight update process.
subsample	0.7	Enhancing the model's generalization capability (using a subset of data).
colsample_bytree	0.7	Improving the model's generalization capability (using a subset of features).
reg_lambda (L2)	1	L2 regularization to reduce model variance.

H. Performance Evaluation

The model's performance evaluation was carried out using the test data, which was excluded from the training phase to objectively assess its generalization capability. Several statistical metrics were employed, including Accuracy, Precision, Recall, F1-score, Confusion Matrix, and the ROC-AUC curve, to provide a comprehensive overview of the model's ability to classify both classes. In addition, a Feature Importance analysis was conducted to understand the relative contribution of each variable to the model's decision-making process.

1) Confusion Matrix: The confusion matrix is used to evaluate the classification model's ability to distinguish between each category in the test data. This 2×2 matrix illustrates the relationship between predicted outcomes and actual conditions, consisting of four main components: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Several key performance metrics are calculated using Equations (6) - (9).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

$$F1 = 2 \times \frac{Precision\ .\ Recall}{Precision + Recall} \tag{9}$$

These metrics collectively provide a quantitative overview of the balance between the model's precision and recall. In the context of bankruptcy prediction, where class imbalance is significant, the interpretation of recall and F1-score become particularly crucial, as both reflect the model's ability to accurately identify bankruptcy cases without being overly influenced by the dominance of the majority class.

2) ROC and AUC Evaluation: The Receiver Operating Characteristic (ROC) curve is employed to assess the classification model's capability to distinguish between positive and negative classes. This curve illustrates the relationship between the True Positive Rate (Recall) and the False Positive Rate (FPR) across various decision thresholds. The Area Under the Curve (AUC) value represents the model's ability to differentiate between bankrupt and non-bankrupt companies the closer the value is to 1, the better the classification performance. Since the dataset is imbalanced, an additional analysis using the Precision-Recall (PR) Curve and the AUC-PR score was conducted to provide a more accurate evaluation of the model's performance on the minority class.

3) *Feature Importance:* In addition to accuracy-based evaluation, a Feature Importance analysis was conducted to measure the relative contribution of each attribute selected through the Chi-Square method in the decision-making process of the XGBoost model trained with data balanced using ROS. In the XGBoost algorithm, the importance level of each feature is internally computed based on gain, which represents the average improvement in model accuracy resulting from data splits made by a particular feature within the decision tree structure. Features with higher importance scores indicate a stronger influence in reducing the model's loss function during the learning process.

This analysis not only serves to explain the model's behavior from a technical perspective (model interpretability) but also carries practical implications in managerial contexts. By identifying the dominant features resulting from the combination of ROS, Chi-Square, and XGBoost, decision-makers can pinpoint the most critical financial factors contributing to bankruptcy risk and design more targeted mitigation strategies.

III. RESULT

A. Results of Data Exploration

The exploratory data results are presented to understand the initial characteristics of the dataset prior to the modelling process. The preliminary inspection of the dataset revealed that there were no missing values across all available attributes. This finding was confirmed through a check of the number of null values per column, which returned a count of zero for every feature indicating that all data entries were complete and no fields were left unfilled.

The distribution of the target variable "Bankrupt?" reveals a significant disparity between the number of non-bankrupt and bankrupt companies. As illustrated in Figure 2, class 0 (non-bankrupt) dominates the dataset with more than 6,000 instances, while class 1 (bankrupt) represents only a small fraction of the total samples. This pronounced imbalance indicates the presence of class imbalance, which may adversely affect the model's ability to accurately learn and identify bankruptcy patterns.

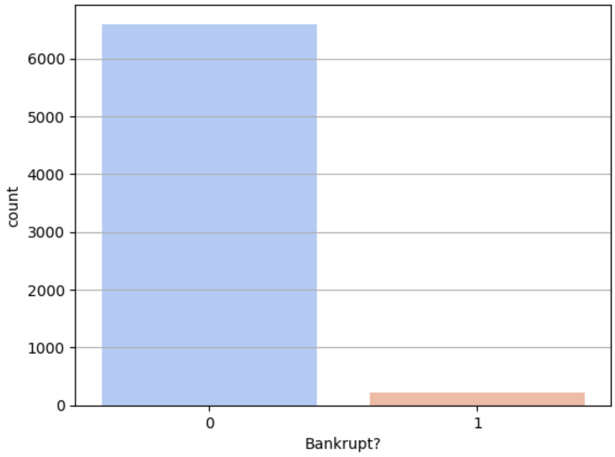

Figure. 2. Target Class Distribution "Bankrupt?"

The evaluation of relationships between numerical features and the target variable was conducted using a correlation heatmap, as shown in Figure 3. From the total of 96 features in the dataset, the fifteen features with the highest correlation values to the "Bankrupt?" label were selected for visualization. This selection aimed to maintain a balance between analytical depth and result readability while focusing the interpretation on the variables that have the most significant influence on the likelihood of bankruptcy.

The results of the heatmap indicate that features such as Net Income to Total Assets, ROA(A) before interest and tax, and ROA(B) before interest and depreciation after tax exhibit relatively strong positive correlations with the target variable, suggesting that profitability plays a crucial role in distinguishing bankrupt from non-bankrupt firms. Conversely, features such as Debt Ratio and Current Liability to Current Assets display strong negative correlations, implying that higher liability ratios tend to increase the likelihood of bankruptcy. Most of the remaining features show low to moderate correlation values, indicating their limited contribution to the overall prediction process.

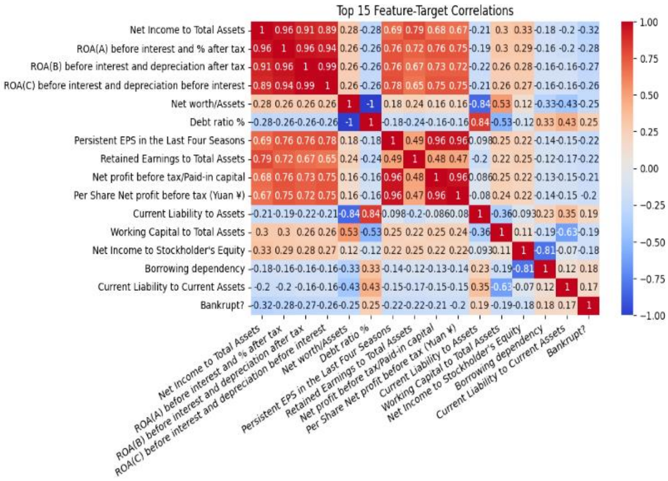

Figure 3. Correlation Heatmap: Top 15 Features and “Bankrupt?” Target

The distribution of the five features with the highest correlation to the “Bankrupt?” variable is illustrated in the histogram shown in Figure 4. The distribution patterns indicate that most features are concentrated within medium to high value ranges, suggesting that the majority of companies in the dataset exhibit relatively stable levels of profitability and financial efficiency.

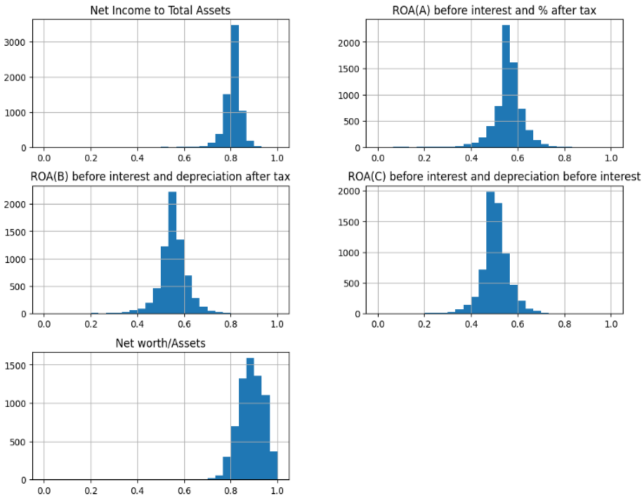

Figure 4. Distribution of the top five features with the highest correlation to the target variable “Bankrupt?”

The features Net Income to Total Assets and Net Worth/Assets exhibit predominantly high-value distributions, indicating that most entities maintain sound financial conditions and are capable of generating proportional profits relative to their assets. Meanwhile, ROA(A), ROA(B), and ROA(C) display greater variability, reflecting differences in operational efficiency across companies that may contribute to bankruptcy risk. These findings emphasize the critical role of profitability variables as early indicators in detecting potential financial distress.

B. *Results of Random Oversampling Implementation*

The class distribution before and after applying ROS is illustrated in Figure 5, showing the change in proportion between the majority class (non-bankrupt) and the minority class (bankrupt). Prior to balancing, the majority class contained 5,279 samples, while the minority class consisted of only 176 samples, yielding an approximate ratio of 30:1. This severe imbalance could potentially reduce the model’s ability to identify bankruptcy patterns accurately due to the dominance of the majority class data.

After applying ROS, the number of samples in the minority class increased to match that of the majority class, reaching 5,279 samples and resulting in a quantitatively balanced distribution. This balance was achieved without altering the original feature distribution structure, thereby preserving the quality of information. With a more proportionate data composition, the model is expected to demonstrate improved sensitivity in detecting bankruptcy cases while simultaneously reducing prediction bias.

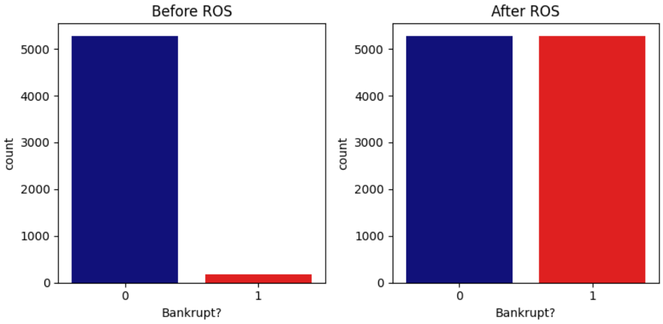

Figure 5. Comparison of target class distribution before and after ROS

To validate the selected oversampling strategy, a performance comparison between ROS and SMOTE was conducted, as summarized in Table III. Both methods achieve comparable overall accuracy. However, differences are observed in minority-class metrics. SMOTE yields higher recall for bankrupt firms, indicating improved detection of minority cases, while ROS demonstrates higher precision and a more stable F1-score. These results highlight distinct trade-offs between sensitivity and prediction reliability across the two oversampling techniques.

TABLE III
ROS AND SMOTE PERFORMANCE COMPARISON

	ROS	SMOTE
Precision (1)	0.45	0.37
Recall (1)	0.41	0.52
F1-Score	0.43	0.43
Accuracy	0.96	0.96

C. *Results of Chi-Square Feature Selection*

The feature selection process produced a ranking of financial variables based on their statistical association with the target variable, derived through the Chi-Square testing method. As illustrated in Figure 6, the visualization highlights the top twenty features according to their Chi-Square scores, indicating the relative significance of each variable in influencing the bankruptcy prediction outcome.

The Chi-Square computation results reveal that the Tax Rate (A) variable achieved the highest ranking, possessing the greatest Chi-Square score, followed by Fixed Assets Turnover Frequency and Cash/Total Assets. This finding indicates that tax burden, efficiency in utilizing fixed assets, and liquidity level are the primary indicators distinguishing financially stable companies from those at risk of bankruptcy. Meanwhile, ratios such as Debt Ratio %, Liability-Assets Flag, and Quick Assets/Total Assets also demonstrate notable influence, as they reflect a company's capital structure, leverage level, and its ability to meet short-term financial obligations.

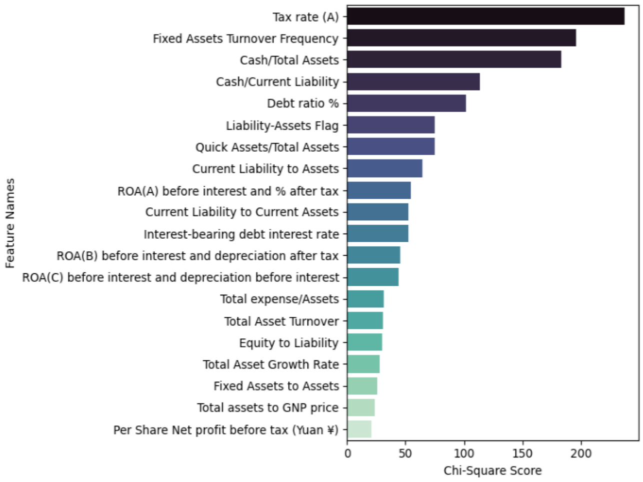

Figure 6. Top twenty financial features based on Chi-Square values for the "Bankrupt?" target variable

Overall, the findings indicate that variables associated with profitability, asset efficiency, and financial structure play a dominant role in shaping the bankruptcy prediction model. Therefore, only features with high Chi-Square values were retained for the subsequent modelling stage to enhance computational efficiency and minimize the potential risk of overfitting.

D. Model Performance Evaluation

As shown in Table IV, the XGBoost model demonstrates excellent performance in predicting the financial condition of companies. The obtained accuracy of 0.9648 indicates that approximately 96% of the test samples were correctly classified. For the majority class (non-bankrupt), the model achieved a precision of 0.9804 and a recall of 0.9833, signifying that it was able to accurately identify financially stable companies with a very low rate of misclassification.

On the other hand, the minority class (bankrupt) achieved a precision of 0.4500 and a recall of 0.4091. Although these values are relatively lower than those of the majority class, they represent a notable improvement compared to models trained without data balancing, which often fail to identify instances belonging to the underrepresented class. The F1-score of 0.4286 further indicates a reasonable trade-off between the model's ability to correctly recognize and accurately predict bankruptcy cases.

Meanwhile, the macro average value of 0.7052 and the weighted average of 0.9640 indicate that the model demonstrates overall predictive stability across both classes. Consequently, the integration of ROS, Chi-Square, and XGBoost methods proves effective in producing a robust classification model that is sufficiently sensitive to the minority class, while remaining relevant for implementation as a decision-support system for early bankruptcy detection.

TABLE IV
XGBOOST MODEL PERFORMANCE RESULTS

	Precision	Recall	F1-Score	Support
0	0.9804	0.9833	0.9818	1320
1	0.4500	0.4091	0.4286	44
accuracy			0.9648	1364
macro avg	0.7152	0.6962	0.7052	1364
weighted avg	0.9633	0.9648	0.9640	1364

To validate the claim of superior performance, the proposed XGBoost model combined with ROS and Chi-Square was compared against a Random Forest baseline using the same preprocessing pipeline. As shown in Table V. The experimental results show that XGBoost achieves higher overall accuracy 0.9648 compared to Random Forest 0.9384. More importantly, for the minority class (bankrupt companies), XGBoost provides substantially higher precision (0.45 vs. 0.31), indicating a lower false positive rate, while maintaining a comparable F1-score. Although Random Forest attains a higher recall, this improvement is accompanied by a significant reduction in precision, suggesting less reliable classification for practical deployment.

TABLE V
XGBOOST AND RANDOM FOREST MODEL PERFORMANCE RESULTS

	XGBoost	Random Forest
Precision (1)	0.4500	0.3077
Recall (1)	0.4091	0.7273
F1-Score	0.4286	0.4324
Accuracy	0.9648	0.9384

At the model performance evaluation stage, a Confusion Matrix analysis was conducted to assess the capability of the XGBoost algorithm in accurately classifying a company's financial condition. Based on the visualization shown in Figure 7, the model demonstrated excellent performance in identifying non-bankrupt firms, correctly classifying 1.298 out of 1.320 actual samples in this category. This indicates that the model achieved a high level of accuracy in recognizing financially stable entities. Although a small number of misclassifications occurred 22 non-bankrupt companies predicted as bankrupt and 26 bankrupt firms classified as non-bankrupt the results still reflect a well-balanced performance between the detection of majority and minority classes. Overall, these findings confirm that the XGBoost algorithm provides reliable and consistent predictive performance, making it a promising decision-support tool for early financial warning systems aimed at identifying potential corporate bankruptcy.

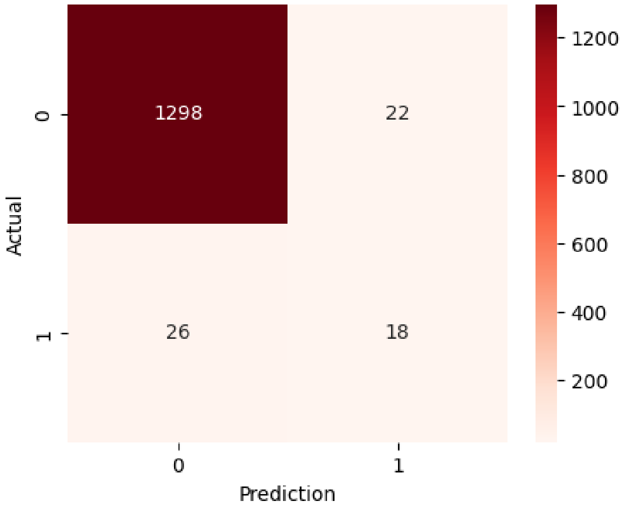

Figure 7. Confusion Matrix

Following the evaluation using the Confusion Matrix, the model's performance was further examined through the ROC curve and the AUC metric.

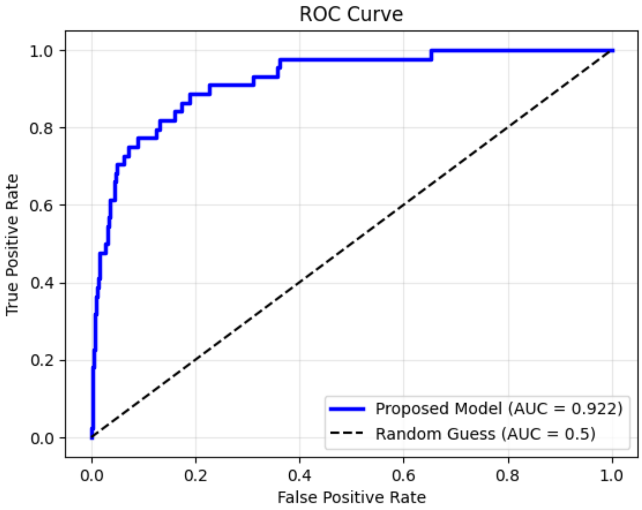

Figure 8. ROC and AUC Curve

As presented in Figure 8, the XGBoost model achieved an AUC value of 0.922, indicating an excellent capability in distinguishing between bankrupt and non-bankrupt firms. An AUC value approaching 1 reflects a high level of class discrimination, where the True Positive Rate consistently exceeds the False Positive Rate across various decision thresholds. The curve's sharp ascent toward the upper left corner of the graph illustrates that the model maintains strong predictive accuracy and stability under different classification scenarios. Hence, these findings confirm that the XGBoost algorithm not only achieves high overall accuracy but is also effective in proportionally mapping bankruptcy probabilities establishing it as a robust and reliable model for corporate bankruptcy prediction.

E. Error Analysis of Classification Results

Based on the confusion matrix in Figure 7, the XGBoost model correctly classified 1,303 out of 1,320 non-bankrupt companies, while 17 cases were incorrectly identified as bankrupt (false positives). For the bankrupt class, only 18 out of 44 cases were correctly predicted, whereas 26 cases were misclassified as non-bankrupt (false negatives). These errors are primarily attributed to data imbalance, as the number of non-bankrupt companies is significantly larger, causing the model to be biased toward recognizing majority class patterns. Additionally, some financial ratios for bankrupt and non-bankrupt companies overlap, making it challenging for the model to distinguish between the two classes. Nevertheless, the true positive count of 18 out of 44 cases indicates that the model still retains the capability to identify a substantial portion of significant bankruptcy patterns. Overall, these results suggest that although the model faces challenges in detecting the minority class, its performance demonstrates a reasonable balance between overall accuracy and sensitivity to bankruptcy cases.

F. Results of Feature Contribution

In addition to evaluating overall performance using the AUC value, an analysis was conducted to assess the contribution of each feature to the model's predictions through a Feature Importance visualization.

Based on the visualization in Figure 9, the features Per Share Net Profit Before Tax, Debt Ratio %, and ROA(B) Before Interest and Depreciation After Tax occupy the highest levels of importance for the model's predictions. This indicates that per-share profitability, debt ratio, and asset management efficiency after depreciation and interest are the most critical factors in distinguishing between bankrupt and non-bankrupt companies. Additionally, variables such as Cash/Current Liability, ROA(C) Before Interest and Depreciation Before Interest, and Current Liability to Current Assets also contribute meaningfully, reflecting that short-term liquidity and the balance between assets and liabilities influence the company's financial stability. Meanwhile, features with low influence, such as Liability-Assets Flag and Quick Assets/Total Assets, suggest that not all financial ratios provide strong signals regarding bankruptcy risk. Overall, these results emphasize that profitability, operational efficiency, and liability management are the primary components determining the accuracy of bankruptcy prediction in the XGBoost model.

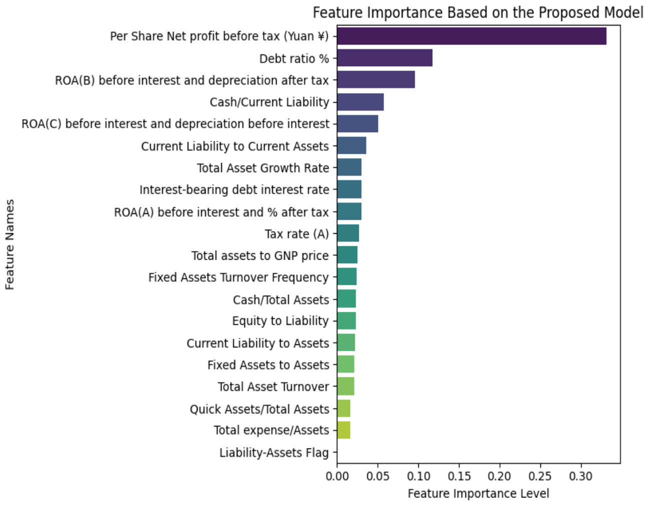

Figure 9. Visualization of Feature Importance XGBoost + ROS + Chi-square

IV. DISCUSSION

A. Interpretation of Results

Data imbalance, where the number of healthy companies far exceeds that of bankrupt ones, poses a major challenge in bankruptcy classification because the model tends to learn patterns from the majority class while neglecting the minority class. In this research, the application of ROS successfully balanced the proportions between the two classes, giving the model an equal opportunity to learn the characteristics of the minority class. As a result, sensitivity to at-risk companies increased, reflected by a rise in recall without compromising the overall accuracy, which remained high. Therefore, the data balancing mechanism is proven to be a crucial preprocessing step in enhancing prediction quality.

The potential risk of overfitting introduced by oversampling was evaluated through a comparative analysis between ROS and SMOTE. The results show no significant degradation in generalization performance on unseen test data, as indicated by stable accuracy and F1-score values. While SMOTE improves recall for the bankrupt class, this gain is accompanied by a decrease in precision, reflecting a higher rate of false positive predictions. In contrast, ROS demonstrates a more balanced trade-off between precision and recall, leading to more consistent minority-class performance. This balance is particularly important in bankruptcy prediction, where excessive false alarms may trigger unnecessary mitigation actions for financially healthy firms. In addition, by replicating existing samples rather than generating synthetic ones, ROS preserves the original feature distribution, which helps limit overfitting when combined with appropriate regularization in the XGBoost framework. Therefore, ROS is selected as the primary oversampling strategy in this research.

Feature selection using the Chi-Square test allows for filtering financial variables so that only the most relevant ones are included in the model. The result is 20 variables with significant influence, such as Tax Rate (A), Fixed Assets Turnover Frequency, and Cash/Total Assets. By eliminating less informative or redundant features, the model becomes more efficient and stable, accelerating training, reducing the risk of overfitting, and enhancing interpretability. The feature importance visualization supports this finding by showing that capital structure and liquidity features carry the greatest influence, confirming that feature selection not only aids computational efficiency but also strengthens the model's theoretical foundation.

Although the proposed model achieves high overall accuracy (0.9648), the relatively low F1-score highlights the impact of class imbalance in bankruptcy prediction. Accuracy is mainly influenced by the majority (non-bankrupt) class and may therefore overstate true performance. In contrast, the F1-score, which balances precision and recall, provides a more appropriate measure of the model's ability to detect bankrupt firms. This discrepancy indicates that bankruptcy detection remains challenging and should be evaluated using class-specific metrics rather than accuracy alone.

The feature importance analysis indicates that profitability-related ratios, particularly Net Income to Total Assets and various forms of Return on Assets (ROA), play a dominant role in the model's decision-making process. This finding is consistent with financial distress theory, which emphasizes that a firm's ability to generate earnings from its asset base is a fundamental determinant of business sustainability. Declining profitability reflects reduced operational efficiency and limited internal financing capacity, thereby increasing reliance on external debt and elevating bankruptcy risk. In addition, leverage and liquidity indicators such as Debt Ratio, Current Liability to Current Assets, and Net Worth/Assets also exhibit substantial importance, aligning with capital structure and liquidity risk theories that associate high financial obligations and short-term payment pressure with an increased likelihood of financial failure. The prominence of these variables suggests that the proposed XGBoost-based model captures economically meaningful signals rather than purely statistical patterns. Overall, the consistency between the identified key features and established financial theory supports the interpretability and theoretical validity of the model, reinforcing its suitability as a reliable decision-support tool for bankruptcy risk assessment under imbalanced financial data conditions.

Furthermore, the XGBoost model was selected due to its adaptive nature and its ability to perform gradient boosting on previous prediction errors. Compared to other models such as Random Forest or MLP, XGBoost can leverage the structure of financial numerical data more efficiently through regularization processes that mitigate overfitting. Consequently, this model provides an optimal balance between prediction accuracy and stability.

The comparative results indicate that XGBoost combined with ROS and Chi-Square provides more reliable performance than Random Forest under the same experimental settings. Although Random Forest achieves higher recall for the minority (bankrupt) class, this improvement is accompanied by a substantial drop in precision, indicating a higher rate of false positive predictions. In contrast, XGBoost demonstrates superior overall accuracy and a better balance between precision and recall, resulting in more stable minority-class performance. This trade-off is particularly important in bankruptcy prediction, where excessive false alarms may reduce the practical usefulness of the model. Therefore, the results support the selection of XGBoost as the primary classifier due to its stronger generalization ability and more consistent classification behavior on imbalanced financial data.

The combination of ROS, Chi-Square feature selection, and adaptive learning (XGBoost) enables this pipeline to capture complex financial patterns while maintaining model stability. The results indicate that this integrative approach not only improves accuracy but also enhances recall and F1-score for the minority class compared to previous studies.

Although the proposed model shows strong performance on the Taiwanese Bankruptcy dataset, its generalizability may be affected by variations in economic conditions, regulatory frameworks, and corporate structures across different contexts. Financial ratios are inherently sensitive to macroeconomic environments and accounting standards, which may influence their predictive effectiveness when applied to other regions or industries. Nevertheless, the proposed framework integrating class balancing, feature selection, and ensemble learning remains adaptable and can be effectively transferred to other datasets through appropriate retraining and calibration.

B. Comparison with Previous Research

As a step to strengthen the validity of this research's results, a comparative analysis was conducted on several previous studies relevant to corporate bankruptcy prediction. This analysis aims to evaluate the effectiveness of the combined methods used ROS, Chi-Square, and XGBoost in enhancing classification performance compared to other models developed earlier. Table VI presents a summary of three previous researches that employed different approaches in terms of dataset, model, data balancing, feature selection, and achieved accuracy. These three studies were selected because they represent the development of machine learning-based bankruptcy prediction techniques in the context of imbalanced financial data.

TABLE VI
COMPARISON WITH PREVIOUS RESEARCH

Dataset	Model	Data Balancing	Feature Selection	Accuracy Results
Taiwanese company data (6,819 records, 96 variables). [27]	MLP	x	x	> 80 %
Annual financial data of Indonesian public companies. [15]	Neural Network (NN)	x	x	86,7 %
IDX-listed company data (2013-2022). [28]	Random Forest	SMOTE	Financial Ratio–Based Feature Selection (manual/domain-based)	96 %
This research	XGBoost	ROS	Chi-Square	96,48%

Based on Table VI, it can be seen that this research achieves higher accuracy and more consistent model performance compared to the three previous studies. Thus, the combination of ROS and Chi-Square methods in this research has proven to enhance the XGBoost model's ability to recognize patterns in the minority class without compromising overall accuracy. These results indicate that the addition of appropriate preprocessing steps contributes significantly to the effectiveness of machine learning-based bankruptcy prediction systems.

V. CONCLUSION

This research successfully developed a corporate bankruptcy prediction model by combining ROS, Chi-Square, and the XGBoost algorithm. Testing results indicate that the model achieved an accuracy of 96% while demonstrating strong capability in recognizing the minority class. The data balancing process through ROS enhanced the model's sensitivity to bankruptcy cases, while Chi-Square feature selection helped identify the most relevant financial variables, such as Tax Rate (A) and Cash/Total Assets. Therefore, this approach is effective in improving prediction performance and can serve as a foundation for developing an early warning system for corporate bankruptcy. From a practical implementation perspective, prediction errors must be carefully managed, particularly false negative cases where financially distressed firms are incorrectly classified as healthy. Such misclassifications may lead to delayed intervention and potentially significant financial losses for investors, creditors, and regulators. Consequently, the proposed model should be employed as a decision-support tool rather than a standalone decision-making system, complementing expert judgment and regulatory assessment. Moreover, real-world deployment requires continuous monitoring, periodic retraining, and contextual adaptation to mitigate operational risks and maintain reliability.

For future research, it is recommended to expand the data scope by integrating financial, macroeconomic, and non-financial indicators to achieve a more comprehensive representation of companies. Hybrid ensemble approaches, such as XGBoost LSTM or Autoencoder Boosting, could be employed to handle the complexity of dynamic data. Additionally, incorporating model interpretability is expected to make predictions more accurate, transparent, and valuable as a decision support system for company management and regulators.

REFERENCES

[1] D. Hafizah, L. Sa, and U. K. A Wahab Hasbullah, “Analisis Komparatif Prediksi Kebangkrutan dengan Metode Altman Z-Score dan Zmijewski X-Score,” *Creative Research Management Journal*, vol. 7, no. 2, p. 127, Dec. 2024, doi: https://doi.org/10.32663/09awrs60.

[2] Heri Triyono and Nurmala Ahmar, “Meta Analisis Hasil Prediksi Kegagalan Perusahaan dengan Pendekatan 4 Model Prediksi Kebangkrutan,” *Journal of Accounting and Finance Management*, vol. 5, no. 3, Aug. 2024, doi: https://doi.org/10.38035/jafm.v5i3.623.

[3] R. Nida, “The impact of digital transformation on financial inclusion: Evidence from MSMEs in Indonesia,” *Jurnal Perspektif Pembiayaan dan Pembangunan Daerah*, vol. 12, no. 4, pp. 2355–8520, Oct. 2024, doi: 10.22437/ppd.v12i4.36399.

[4] D. Saputra, A. M. Yudha, and T. Ulnisa, “Pengaruh Tingkat Suku Bunga, Nilai Tukar Dan Inflasi Terhadap Nilai Perusahaan Dengan Profitabilitas Sebagai Variabel Moderasi Pada Perusahaan Property Dan Real Estate Yang Terdaftar Di Bei 2017-2021,” *JAF- Journal of Accounting and Finance*, vol. 8, no. 1, p. 54, Mar. 2024, doi: 10.25124/jaf.v8i1.7224.

[5] A. E. Prihatini and D. Purbawati, “Analisis kesehatan Keuangan dengan Menggunakan Metode Altman Z-Score Pada PT Tiga Pilar Sejahtera Food Tbk,” *Jurnal Administrasi Bisnis*, vol. 10, no. 2, pp. 155–164, Sep. 2021, doi: 10.14710/jab.v10i2.36791.

[6] M. R. Dewi and D. Susilaningrum, “A Hybrid Model to Enhance The Performance of Classifier in Financial Distress Prediction,” *Indonesian Journal of Applied Informatics*, vol. 9, no. 1, p. 138, Nov. 2024, doi: 10.20961/ijai.v9i1.94725.

[7] R. Rinofah, R. Kusumawardhani, and V. A. Maha Putri, “Factors Affecting Potential Company Bankruptcy During The Covid-19 Pandemic,” *Jurnal Keuangan dan Perbankan*, vol. 26, no. 1, pp. 208–228, Mar. 2022, doi: 10.26905/jkdp.v26i1.6752.

[8] A. Hartono, W. R. Dita, and I. F. Ulfah, “Analysis of the Altman, Springate, Zmijewski, and Grover Methods in Predicting Bankruptcy in Retail Electronics Sub Sector Companies Listed on the Indonesia Stock Exchange for the 2019-2022 Period,” *Ekuilibrium : Jurnal Ilmiah Bidang Ilmu Ekonomi*, vol. 20, no. 2, pp. 341–353, Sep. 2025, doi: 10.24269/ekuilibrium.v20i2.2025.pp341-353.

[9] Y. Nurhayati and E. F. Komara, “Predictive Analysis of Financial Distress Using the Altman Z-Score Method on Companies in the Trade, Service & Investment Sector Listed on the Indonesia Stock Exchange in 2019-2023,” *Formosa Journal of Applied Sciences*, vol. 4, no. 8, pp. 2531–2546, Aug. 2025, doi: 10.55927/fjas.v4i8.299.

[10] A. Kurani, P. Doshi, A. Vakharia, and M. Shah, “A Comprehensive Comparative Study of Artificial Neural Network (ANN) and Support Vector Machines (SVM) on Stock Forecasting,” *Annals of Data Science*, vol. 10, no. 1, pp. 183–208, Feb. 2023, doi: 10.1007/s40745-021-00344-x.

[11] F. M. Irvan, “Comparative Analysis Of Machine Learning and Deep Learning Models Integrated With Altman Z-Score For Financial Distress Prediction In Companies Listed On The Indonesia Stock Exchange (IDX),” *EKOMBIS REVIEW: Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 12, no. 2, Apr. 2024, doi: 10.37676/ekombis.v12i2.5478.

[12] R. Saputra, S. Sunardiyo, A. Nugroho, and S. Subiyanto, “Analisis Arsitektur Jaringan Syaraf Tiruan-Multilayer Perceptron untuk Efektivitas Estimasi Beban Energi Listrik PT. PLN (Persero) UP3 Salatiga,” *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika*, vol. 11, no. 3, p. 664, Jul. 2023, doi: 10.26760/elkomika.v11i3.664.

[13] W. I. Sabilla and C. Bella Vista, “Implementasi SMOTE dan Under Sampling pada Imbalanced Dataset untuk Prediksi Kebangkrutan Perusahaan,” *Jurnal Komputer Terapan*, vol. 7, no. 2, pp. 329–339, Dec. 2021, doi: 10.35143/jkt.v7i2.5027.

[14] D. J. Maulana, 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)*, vol. 8, no. 1, pp. 54–61, Feb. 2024, doi: 10.29207/resti.v8i1.5140.

[15] D. S. Rahayu, H. Suhartanto, and A. Husodo, “Assessing Data Imbalance in Financial Distress Prediction: A Comparative Approach of Machine Learning and Economic Models,” *JOIV : International Journal on Informatics Visualization*, vol. 9, no. 5, pp. 1929–1941, Sep. 2025, doi: http://dx.doi.org/10.62527/joiv.9.5.3397.

[16] B. Siswoyo, Z. Abal Abas, A. N. Che Pee, R. Komalasari, and N. Suryana, “Ensemble machine learning algorithm optimization of bankruptcy prediction of bank,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, p. 679, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp679-686.

[17] D. Nurmalasari, H. R. Yuliantoro, and D. H. Qudsi, “Improving Panic Disorder Classification Using SMOTE and Random Forest,” *Journal of Applied Informatics and Computing*, vol. 8, no. 2, pp. 272–279, Oct. 2024, doi: 10.30871/jaic.v8i2.8315.

[18] T. Kurniawan, L. Hermawanti, and A. N. Safriandono, “Interpretable Machine Learning with SHAP and XGBoost for Lung Cancer Prediction Insights,” *Journal of Applied Informatics and Computing (JAIC)*, vol. 8, no. 2, p. 296, Dec. 2024, doi: https://doi.org/10.30871/jaic.v8i2.8395.

[19] I. K. Ananda, A. Z. Fanani, D. Setiawan, and D. F. Wicaksono, “Penerapan Random Oversampling dan Algoritma Boosting untuk Memprediksi Kualitas Buah Jeruk,” *Edumatic: Jurnal Pendidikan Informatika*, vol. 8, no. 1, pp. 282–289, Jun. 2024, doi: 10.29408/edumatic.v8i1.25836.

[20] A. T. P. Subandono and D. Ariatmanto, “Optimizing Feature Selection in Sentiment Analysis of Bank Saqu: A Comparative Study of SVM and Random Forest using Information Gain and Chi-Square,” *SISTEMASI*, vol. 14, no. 3, p. 1205, May 2025, doi: 10.32520/stmsi.v14i3.5106.

[21] D. Kurnia, M. Itqan Mazdadi, D. Kartini, R. Adi Nugroho, and F. Abadi, “Seleksi Fitur dengan Particle Swarm Optimization pada Klasifikasi Penyakit Parkinson Menggunakan XGBoost,” *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 5, pp. 1083–1094, Oct. 2023, doi: 10.25126/jtiik.2023107252.

[22] E. Hokijuliandy, H. Napitupulu, and F. Firdaniza, “Analisis Sentimen Menggunakan Metode Klasifikasi Support Vector Machine (SVM) dan Seleksi Fitur Chi-Square,” *SisInfo : Jurnal Sistem Informasi dan Informatika*, vol. 5, no. 2, pp. 40–49, Aug. 2023, doi: 10.37278/sisinfo.v5i2.670.

[23] E. Mustika Sari, C. Sabila, R. Fakhrizal Adam, and R. Kurniawan, “Analisis dan Prediksi Indeks Kualitas Udara Jakarta: Penerapan Algoritma XGBoost,” *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 11, no. 2, pp. 161–169, Sep. 2025, doi: 10.25077/TEKNOSI.v11i2.2025.161-169.

[24] R. E. Ako *et al.*, “Effects of Data Resampling on Predicting Customer Churn via a Comparative Tree-based Random Forest and XGBoost,” *Journal of Computing Theories and Applications*, vol. 2, no. 1, pp. 86–101, Jun. 2024, doi: 10.62411/jcta.10562.

[25] G. Airlangga, “Comparative Study of XGBoost, Random Forest, and Logistic Regression Models for Predicting Customer Interest in Vehicle Insurance,” *sinkron*, vol. 8, no. 4, pp. 2542–2549, Oct. 2024, doi: 10.33395/sinkron.v8i4.14194.

[26] R. Andespa, K. Sadik, C. Suhaeni, and A. M. Soleh, “Evaluating Random Forest and XGBoost for Bank Customer Churn Prediction on Imbalanced Data Using SMOTE and SMOTE-ENN,” *MEDIA STATISTIKA*, vol. 18, no. 1, pp. 25–36, Oct. 2025, doi: 10.14710/medstat.18.1.25-36.

[27] R. F. Brenes, A. Johannssen, and N. Chukhrova, “An intelligent bankruptcy prediction model using a multilayer perceptron,” *Intelligent Systems with Applications*, vol. 16, Nov. 2022, doi: 10.1016/j.iswa.2022.200136.

[28] F. T. Kristanti, M. Y. Febrianta, D. F. Salim, H. A. Riyadh, and B. A. H. Beshr, “Predicting Financial Distress in Indonesian Companies using Machine Learning,” *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 17644–17649, Dec. 2024, doi: 10.48084/etasr.8520.

[29] “Taiwanese Bankruptcy Prediction,” UCI Machine Learning Repository. Accessed: Oct. 30, 2025. [Online]. Available: https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction