Advances in Artificial Intelligence and Machine Learning; Research 5 (2) 3663-3680 Received 28-01-2025; Accepted 19-04-2025; Published 26-04-2025

Improving Financial Distress Prediction through Clustered SMOTE for Imbalanced Data

Kalina Kitova
Sofia University 'St. Kliment Ohridski',
Faculty of Economics and Business Administration,
Bulgaria

Borislava Toleva vrigazova@uni-sofia.bg
Sofia University 'St. Kliment Ohridski',
Faculty of Economics and Business Administration,
Bulgaria

Ivan Ivanov
Sofia University 'St. Kliment Ohridski',
Faculty of Economics and Business Administration,
Bulgaria

Corresponding Author: Borislava Toleva

Abstract

Financial distress prediction remains fundamental to identifying troubled businesses since it determines business stability along with economic forecast accuracy. The research evaluates the Synthetic Minority Over-sampling Technique (SMOTE) to correct class imbalance issues in financial distress prediction by studying its results when standardized through clustering and non-clustering approaches. The research determines how K-means clustering strengthens SMOTE by applying data balancing techniques inside separate clusters to improve model predictions for financial distress. Combining K-means clustering with SMOTE substantially improves model performance because XGBoost demonstrates the peak results, including 99% accuracy and 99% F1 score. Incorporating clustering methods helps SMOTE produce more accurate synthetic samples, achieving better predictive accuracy by improving class balance. According to these results, combining clustering methods and SMOTE demonstrates great potential for financial distress prediction in imbalanced datasets.

1. INTRODUCTION

Business and economic stability depends heavily on accurate forecasting of financial distress in today's fast-changing economic environment. Financial distress prediction processes help stakeholders evaluate company bankruptcy chances to implement safety measures that reduce possible hazards. Predicting financial distress proves difficult because such systems usually contain datasets with an unusually uneven distribution of classes. The outnumbers of healthy companies within

Citation: Kalina Kitova, et al. Improving Financial Distress Prediction through Clustered SMOTE for Imbalanced Data. Advances in Artificial Intelligence and Machine Learning. 2025;5(2):208.

financial datasets make predicting distress particularly difficult [1]. Traditional machine learning tools experience difficulty making accurate predictions about financially distressed companies because they belong most often to the minority class [2].

Advanced methods must be employed since the existing imbalance needs attention to boost the performance of prediction models. Numerous machine learning and data mining techniques have evolved to identify financial distress, ranging from ensemble models and deep learning approaches to more traditional models such as decision trees and logistic regression. Scientists in this research domain apply ensemble learning methodologies and synthetic data creation strategies combined with feature selection methods to enhance model performance. Research demonstrates that SMOTE is commonly used to generate synthetic samples for the minority class ([3, 4]). Researchers have studied different methods that unite clustering methods with SMOTE to create more effective synthetic samples while preserving data structures ([5, 6]). However, the progress made by human language technology has not yet resolved key obstacles to its adoption. The traditional machine learning methods, including Decision Trees (DT), Support Vector Machines (SVM), and Logistic Regression (LR), produce poor predictive results and computational efficiency when analyzing extensive data and skewed class distributions [7]. Most models fail to provide accurate predictions because they ignore what data structure patterns exist in specific data clusters. The urgent need for new advanced methods that address these issues has emerged in response to these problems.

The aim of this study is to fill this gap. The research objective focuses on enhancing SMOTE prediction capabilities for financial distress in imbalanced data by implementing K-means clustering. The proposed method divides data into smaller clusters of homogeneous properties for better sample synthesis, thus decreasing data variability. The method arranges data better using its strategic approach to handle data organization and class unbalanced patterns to improve predictive results. The proposed methodology efficiently distinguishes distressed companies from healthy ones and boosts prediction model accuracy, precision levels, and recall performance. K-means clustering with SMOTE enables the development of stronger transferable models for predicted data. This approach shows enhanced performance results in XGBoost, RF, SVM, DT, and LR analyses.

Financial distress prediction performance becomes the main focus of our proposed methodology, which utilizes K-means clustering and SMOTE to optimize class imbalance treatment and data structure utilization. The proposed clustering method seeks to boost SMOTE effectiveness by dividing datasets into smaller clusters that contain data with similar properties to improve the quality of synthesized samples and minimize the data noise from high variance. The K-means clustering technique, alongside various other clustering methods, has widespread adoption across multiple application areas because of its ability to detect intrinsic data patterns for more accurate analysis, according to [8, 9]. K-means clustering splits data points into separate clusters according to their similarity levels, after which SMOTE generates new synthetic samples that match the distinct characteristics of each cluster. The strategy helps rectify class imbalance problems by enhancing minority class presence in each cluster to achieve better prediction results.

Clustering procedures in the SMOTE process deliver multiple positive results. Clustering allows researchers to spot implicit patterns of natural data groupings that would otherwise remain undetected if viewing the whole dataset at once [10]. When SMOTE operates within separate clusters, it creates synthetic data points that maintain relations with actual data patterns, which results in fewer

implausible sample appearances [11, 12]. Local adjustments of the dataset through this approach result in improved synthetic sample quality, leading to superior prediction outcomes.

The primary classification function relies on XGBoost (Extreme Gradient Boosting), our main machine learning model, because it provides efficient balance management for imbalanced datasets and exceptional predictive accuracy [13]. XGBoost's ability to process non-linear feature relationships along with interactions results in superior results across different domains, such as financial distress prediction [14, 15]. Our main objective is to enhance model discrimination between distressed and healthy companies through a K-means clustering and SMOTE combination, boosting precision-level, recall-level, and predictive accuracy.

In addition to XGBoost, we evaluate our approach using several other classification models, including Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression, to assess the generalizability and robustness of the proposed methodology. While these models have been successfully applied to financial distress prediction in previous studies [16–18], our results suggest that the combination of K-means clustering and SMOTE leads to significant improvements in performance, particularly with XGBoost. In particular, including clustering facilitates better handling of class imbalance, leading to more balanced predictions across all models, with XGBoost achieving the highest accuracy and F1 score.

According to our research findings, Cluster analysis and SMOTE prove to be an innovative and effective method for financial distress prediction. Addressing class imbalance and data structural methods improves predictive accuracy through techniques that surpass traditional approach performance. The research results create vital implications for financial analysts, risk managers, and decision-makers who count on precise predictions of financial distress to confront economic uncertainty and make better business plans. This approach helps researchers develop sophisticated machine-learning methods for financial forecasting. It also guides additional work on combining clustering techniques with SMOTE and other data improvement techniques in financial distress modeling.

The study introduces an innovative financial distress prediction enhancement framework combining K-means clustering and SMOTE. The study results show that clustering brings better results to synthetic sample production and boosts financial distress prediction accuracy. The findings of this investigation advance financial distress prediction research in combination with machine learning models and data preprocessing methods for managing unbalanced datasets.

1.1 Key Contributions of the Study

The key contributions of this study are as follows:

- A novel solution for predicting financial distress is proposed, combining SMOTE with clustering and exploring how clustering improves the model's performance.
- SMOTE generates synthetic examples, and clustering (via K-means) addresses class imbalance and improves prediction accuracy.

- A comparative analysis evaluates the differences between models using only SMOTE and those combining clustering with SMOTE, assessing their impact on prediction accuracy, class balance, and the adaptability of synthetic samples to the real data structure.
- Several machine learning models (XGBoost, Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression) are evaluated to assess the impact of applying SMOTE and clustering on their performance.
- Experimental tests demonstrate that the combination of SMOTE and clustering significantly improves prediction accuracy, with XGBoost achieving the best results.

1.2 Structure of the Paper

The paper is structured as follows: Section 2 reviews related research in the fields of financial distress prediction and class imbalance methods. Section 3 describes the study's methodology, including the approaches for SMOTE and clustering and the models used. Section 4 presents the experimental results, including an analysis of the performance of different models and the combination of SMOTE and clustering. Section 5 concludes the study and offers recommendations for future research.

2. RELATED WORKS

Research on financial distress prediction remains essential because stakeholders need it to address upcoming risks that threaten the company and economic stability. The development of financial distress prediction models benefited from various proposed methods that focused on solving problems related to class imbalance and high-dimensional data while improving computational efficiency.

Discriminant analysis and logistic regression formed the basis for financial distress prediction according to traditional investigative methods [19]. The rise of sophisticated datasets made machine learning techniques more popular because they handle detailed data sources and effectively create non-linear correlations. The popular data classification methods are SVM and DT, where SVM demonstrates exceptional performance in achieving high accuracy [20]. The success of these predictive models faces a limitation when applied to class-imbalanced financial distress prediction datasets because healthy companies significantly outnumber distressed companies.

Recent advancements have introduced more sophisticated techniques to address these challenges. One such method is SMOTE, widely applied to balance imbalanced datasets by generating synthetic samples for the minority class [3]. Additionally, newer approaches such as cluster-based instance selection (CBIS) combine clustering analysis with instance selection to improve the performance of classifiers, demonstrating that clustering techniques can also effectively address class imbalance by grouping similar data samples and filtering out unrepresentative ones [21, 22].

Two alternative approaches for prediction enhancement consist of ensemble methods and deep learning models. XGBoost demonstrates exceptional performance in financial distress prediction challenges since it analyzes both dataset imbalance and feature interrelations successfully [23]. The Multi-Layer Perceptron deep learning model enhanced with genetic algorithms successfully boosted

forecasting precision and attribute identification. Multiple forecasting approaches experience difficulties working with extensive datasets and understanding multiple interrelated features [24].

Research in this field now uses new approaches alongside previously described experimental techniques and theoretical methods. In his research, Green Arther Sandag examined the same dataset used in the current study and suggested implementing a Bagging Classifier as an ensemble learning algorithm to detect financial distress [25]. The best performance achieved in the creation of a predictive model for company health was reported with the Bagging Classifier, yielding an Accuracy of 97.01%, Recall of 96.2%, Precision of 97.36%, an RMSE of 0.183, and an F-score of 97.03%. This represents the highest performance achieved in their study.

Al Ali et al. [26] have again analyzed the same dataset used in the current study and developed CWCSGDFL as a data grouping methodology that uses the Chinese Whisper Clustered Stochastic Gradient Descent Federated Learning method to enhance efficiency levels for sample data. The method achieves better predictive accuracy and lower computational complexity without neglecting data imbalance issues. The results revealed powerful performance through 96% Accuracy, 94% Precision, 98% Recall, and 98% F-measure, which exceeded alternative techniques [26].

In comparison, our research develops existing methods by joining SMOTE functions with K-means clustering to generate superior-quality synthetic samples for financial distress forecasting. A performance assessment of XGBoost and Random Forest alongside Support Vector Machine and Decision Tree and Logistic Regression occurs to determine how SMOTE with clustering improves financial distress prediction ability. Our results indicate that the combination of K-means clustering and SMOTE significantly improves model performance, particularly for XGBoost, and offers a robust solution for financial distress prediction. This approach could contribute to developing more accurate and efficient models for forecasting financial distress in imbalanced datasets.

3. PROPOSED METHODOLOGY FOR FINANCIAL DISTRESS PREDICTION THROUGH CLUSTERING AND SMOTE

Many machine learning techniques to predict financial distress are frequently used, yet researchers have not confirmed whichstrategy is the best solution for predicting financial distress [27, 28]. This research proposed a new predictive approach to enhance the prediction accuracy of financial company distress using a clustering method with K-means and SMOTE. Data clustering and classification features are the key operational components of the proposed methodology.

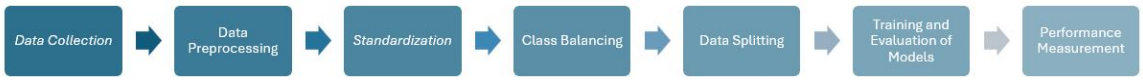

Figure 1: Methotology 1: with SMOTE only

Figure 2: Methotology 2: Clustering prior to SMOTE

3.1 Data Collection.

The dataset used in this approach came from Kaggle's Financial Distress Prediction Dataset [29]. The dataset comprises 86 features along with 3,672 instances that describe multiple companies. The analyzed dataset presents an unbalanced distribution because it includes 136 distressed companies but 3,546 healthy companies. Financial distress is the target variable, and the company should be considered healthy if it is greater than -0.50 (0). Otherwise, it would be regarded as financially distressed (1). The variables from x1 to x83 constitute financial and non-financial characteristics that serve as classification indicators.

3.2 Data Preprocessing.

Data preprocessing is thefirst step before applying any class balancing technique. The target variable, *Financial Distress*, is converted into a binary format. Companies with a financial distress value of ≤ -0.50 are labeled as distressed (1), while all others are categorized as healthy (0). Label encoding transformation applies to the categorical variable "x80" before standardizing all the features to ensure proper scale compatibility for clustering analysis.

3.3 Data Standardization.

StandardScaler is applied to the data to standardize the features, normalize the data, and eliminate scale discrepancies between different features [30].

3.4 Data Clustering.

The analysts perform financial distress data examination as their initial research step. In Methodology 2 (FIGURE 2.), the data evaluation employs the K-means clustering method to detect variations between data points. It requires specifying the number of clusters and scales well to large datasets [31]. The K-means algorithm aims to select centroids that minimize the momentum or sum of squares of distances within clusters, using the following formula:

$$\int_{i=0}^{n} \frac{min}{\mu_{i \in} C} \left(||x_i - \mu_j||^2 \right) \tag{1}$$

Where x_i are the data points, μj are the clusters' centroids, and $||x_i - \mu_j||^2$ is the square of the distance between the point x_i and the centroid μ_j.

The algorithm minimizes this value (eq. 1) to ensure the best separation of the data into different clusters [32].

Silhouette analysis finds the perfect number of clusters to achieve successful data organization. The clustering quality assessment method uses a comparison process that measures internal cluster similarity and differences between clusters [33]. The Silhouette Coefficient compares the average distance of a point to others in the same cluster (a) with the average distance to points in the nearest cluster (b). The formula is:

$$s = \frac{b - a}{\max(a, b)} \tag{2}$$

According to the results, a two-cluster solution provides optimal separation and cohesion of data points, yielding a maximum Silhouette Score. After K-means clustering operates on the data, it forms two distinct groups with 2,550 observations in the first section and 1,122 observations in the second. The dataset segmentation offers insights into data concepts, allowing balancing methods such as SMOTE to function more effectively. Implementing SMOTE following clustering uses specific clustering data to produce synthetic instances that maintain the data structure and optimize model performance [34, 35].

3.5 SMOTE for Handling Imbalanced Data.

The data contains a severely unbalanced distribution because healthy company observations outnumber distressed companies in a 26:1 ratio (Table 1).

Table 1: Dataset description before and after oversampling.

	Before oversampling	**After oversampling**
Healthy	3536	3536
Distressed	136	3536

After clustering, the focus shifts to addressing the data imbalance. SMOTE generates new examples for the minority class by selecting k nearest neighbors for a given sample [36]. A new sample is then created using the following formula:

$$x_{new} = x_i + \lambda \times (x_{zi} - x_i) \tag{3}$$

where λ is a random number between 0 and 1, controlling the distance between the original sample and its neighbor. This creates new synthetic examples connecting the original sample and its nearest neighbor. Each clustered group receives SMOTE treatment to balance the classes before the classification model receives more unbiased training [37]. After applying SMOTE, the observations from both classes are balanced (Table 1). The results become more precise, and the model performance improves because the training contains a balanced representation of financially distressed and healthy companies.

3.6 Data Splitting.

In order to train the model, the data must be split, tested with a subset of the data, and computed with accuracy measures to determine the model's performance in the final stage before applying the machine learning model. Training data and test data were created from the dataset. The training data contained 70% of the total dataset, and the testing data only contained 30% of the complete dataset.

4. PREDICTING FINANCIAL DISTRESS USING A MACHINE LEARNING MODEL

4.1 Algorithm.

Machine learning approaches were used to estimate Financial Distress to meet this aim. The selected classification algorithms demonstrate the accuracy, interpretability and proper management capabilities of unbalanced data sets. We employed the following five machine learning models in our research. Table 2 provides details of the parameters used for each model.

Table 2: Hyperparameters for Each Model

Model	Parameters
LR	LogisticRegression(solver='lbfgs', max_iter=5000)
SVM	SVC(kernel='rbf',probability=True)
DT	DecisionTreeClassifier(random_state=61)
RF	RandomForestClassifier(n_estimators=30, criterion='entropy', max_depth=10, min_samples_leaf=2, random_state=42)
XGBoost	xgb.XGBClassifier(random_state=61)

4.1.1 Logistic regression.

LR is a method for predicting the probability that a given input will fall into category "1". It uses the sigmoid function to perform the analysis [38].

$$g(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

4.1.2 Support vector machine classifier.

SVM is a machine learning method used for classification and prediction of results. It is used for pattern recognition and data segmentation, helping to distinguish and predict the output variable. SVM is popular in areas such as pattern recognition and security penetration testing [39].

$$f(x) = \sum a_j y_j K\left(x_j x\right) + b \tag{5}$$

4.1.3 Decision tree classifier.

A DT is a machine learning method used primarily for categorization. In it, nodes represent data features and paths represent prior information, with each node providing a conclusion based on those features [40].

4.1.4 Random forest classifier.

RF is a prediction model that uses multiple decision trees to optimize performance. Unlike Decision Tree, which can be subject to overfitting, RF applies techniques such as bagging and boosting to improve accuracy and avoid overfitting [41].

$$Gini = 1 - \int_{i=1}^{C} (p_i)^2 \tag{6}$$

4.1.5 Extreme gradient boosting classifier.

XGBoost is a machine learning algorithm used for tasks such as classification that combines predictions from multiple individual models (typically decision trees) in an iterative process. It uses gradient descent optimization to minimize errors and includes overfitting reduction techniques and parallel processing for more efficient computations. [42].

4.2 Measure.

In order to evaluate the performance of the model, following metrics are used.

4.2.1 Accuracy.

Accuracy indicates the proportion of correct predictions (both true positives and true negatives) out of all predictions. It's calculated as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{7}$$

4.2.2 Precision.

Precision measures the proportion of correctly predicted positive cases out of all predicted positives. It's calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

4.2.3 Specificity.

Specificity or True Negative Rate measures how well the model identifies negative cases. Higher specificity indicates better performance in correctly identifying negatives with fewer false positives. It's calculated as:

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

4.2.4 Sensitivity.

Sensitivity or Recall measures the proportion of actual positives correctly identified. It's calculated as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{10}$$

4.2.5 F1 score.

F1 Score is the harmonic mean of precision and recall, providing a balanced performance measure. It's calculated as:

$$F1\ Score = \frac{2x\ Precision\ \times Sensitivity}{Precision\ \times Sensitivity} \tag{11}$$

4.2.6 AUC score.

Area Under the Curve measures the model's ability to distinguish between classes. A higher AUC indicates better performance, with one being perfect and 0.5 being no better than random guessing.

4.2.7 RMSE.

Root Mean Square Error measures the difference between predicted and actual values. Lower RMSE indicates better model accuracy.

4.2.8 Log Loss.

Log Loss quantifies the accuracy of a classifier by penalizing wrong classifications. Lower values indicate better performance..

4.2.9 MCC.

Matthews Correlation Coefficient is a measure of the quality of binary classifications. A value closer to 1 indicates a better model, while -1 indicates a poor model.

4.2.10 Cohen's Kappa.

Cohen's Kappa measures the agreement between predicted and actual values, correcting for chance. A higher value indicates better agreement.

4.2.11 Processing time.

Processing time refers to the model's time to complete the training or prediction process. A shorter processing time is generally preferred for efficiency.

4.2.12 Confusion matrix.

The confusion matrix evaluates a classification model's performance by comparing predicted and actual outcomes, showing true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). These terms help calculate performance metrics like accuracy, precision, and recall [43].

4.3 Results from Previous Studies.

The findings from previous research using this dataset are discussed in this section, along with the work conducted by Sandag and Green Arther [25] and Al Ali and Khedr [26]. Sandag and Green Arther experimented with numerous classification models until they discovered that the Bagging Classifier provided the most effective solution. It reached 97.01% Accuracy with 97.36% Precision and 96.2% Sensitivity, 97.03% F1 Score, and 0.183 RMSE (Table 3). Ensemble-based methods prove effective in predicting financial distress according to the reported experimental outcome. The CWCSGDFL model developed by Al Ali and Khedr [26] represents a combination of clustering and federated learning methods to improve predictive accuracy and handle data imbalance conditions and performance efficiency. The proposed method delivered 97.61% Accuracy as its best result, exceeding multiple competing solutions that included FL, AWOA-DL, MLP-ANN, and CUS-GBDT (Table 4).

Table 3: Performance Metrics of different Models from Sandag and Green Arther

Model	Accuracy	Precision	Sensitivity	F1 Score	RMSE
Bagging Classifier	0.9701	0.9736	0.9620	0.9703	0.1830
SVM	0.9445	0.9750	0.9354	0.9493	0.3410
LR	0.9325	0.9638	0.9052	0.9504	0.4220
DT	0.9450	0.9810	0.9516	0.9589	0.2250

Table 4: Performance Metrics of different Methods with 1,050 Instances from Al Ali & Khedr

Method	Accuracy	Precision	Sensitivity	F1 Score
CWCSGDFL	0.9761	0.9800	0.9900	0.9840
FL Model	0.9000	0.9380	0.9510	0.9440
AWOA-DL	0.8777	0.9320	0.9380	0.9340
MLP-ANN	0.8512	0.8930	0.9320	0.9120
CUS-GBDT	0.8258	0.8750	0.8800	0.8770

4.4 Experimental Results.

In this section, we present the results of the experiments, including the models' baseline performance without any balancing techniques, followed by the performance of two methodologies that incorporate SMOTE.

4.4.1 Baseline results.

Before applying any balancing techniques, we evaluated the models on the original imbalanced dataset to observe how the results would look without addressing the class imbalance. The results of these baseline models are as follows (TABLE 5):

Table 5: Baseline Model Performance

Model	Accuracy	Precision	Specificity	Sensitivity	F1 Score	AUC score	RMSE	Log Loss	MCC	Cohen's Kappa	Processing time (sec)
LR	0.9628	0.5625	0.9934	0.2093	0.3051	0.9263	0.1929	0.1043	0.3281	0.2901	0.10
SVM	0.9637	0.0000	1.0000	0.0000	0.0000	0.8942	0.1905	0.1221	0.0000	0.0000	0.65
DT	0.9483	0.3143	0.9773	0.2500	0.2785	0.6137	0.2274	1.7865	0.2538	0.2520	0.19
RF	0.9628	0.5833	0.9953	0.1628	0.2545	0.8894	0.1929	0.2168	0.2949	0.2416	0.23
XGBoost	0.9646	0.7273	0.9972	0.1818	0.2909	0.9148	0.1881	0.1603	0.3525	0.2794	0.25

Given the highly imbalanced dataset, the models showed high accuracy, but this was mainly due to the dominance of the negative class. XGBoost exhibited the bestoverall performance, with high precision, low RMSE, and superior AUC, indicating its ability to handle the imbalance better. Models like Logistic Regression and Random Forest showed reasonable specificity but struggled with low sensitivity, reflecting difficulty in identifying the minority class. Despite its high specificity, SVM had very low precision and sensitivity, making it less effective for predicting the positive class. The results highlight the need for balancing techniques like SMOTE to improve sensitivity and overall model performance in imbalanced datasets.

4.4.2 Methodology 1: Results with SMOTE application.

In this section, we present the results of the models after applying Methodology 1 (FIGURE 1.), which involves using the SMOTE technique to balance the dataset. This step was performed to

observe the effects of SMOTE alone, and these results will be later compared to the performance achieved by adding clustering to the process, allowing us to assess the impact of clustering on improving SMOTE's effectiveness. The results of these models after applying SMOTE are as follows:

Table 6: Methodology 1: Results with SMOTE Application

Model	Accuracy	Precision	Specificity	Sensitivity	F1 Score	AUC score	RMSE	Log Loss	MCC	Cohen's Kappa	Processing time (sec)
LR	0.9123	0.8901	0.8826	0.9418	0.9152	0.9580	0.2961	0.2450	0.8261	0.8246	0.15
SVM	0.9581	0.9227	0.9160	1.0000	0.9598	0.9902	0.2048	0.0957	0.9194	0.9161	3.11
DT	0.9548	0.9347	0.9343	0.9760	0.9549	0.9551	0.2127	1.5626	0.9104	0.9096	0.47
RF	0.9637	0.9413	0.9354	0.9908	0.9654	0.9956	0.1905	0.1161	0.9286	0.9273	0.62
XGBoost	0.9788	0.9588	0.9581	1.0000	0.9790	0.9985	0.1456	0.0570	0.9585	0.9576	0.66

Applying SMOTE significantly improved model performance compared to the baseline results. Accuracy, precision, and F1 score increased across all models, with XGBoost achieving the highest values (Accuracy: 0.9788, F1: 0.9790). Sensitivity improved notably, reaching 1.0000 for XGBoost and SVM, while AUC scores also increased, indicating better class separation. Lower RMSE and Log Loss reflect more reliable predictions. Overall, SMOTE effectively addressed class imbalance, enhancing the models' ability to classify positive and negative cases correctly.

4.4.3 Methodology 2: Results with clustering prior to SMOTE.

In this section, we present the results of Methodology 2 (FIGURE 2.), where clustering was applied before SMOTE. This approach involved first segmenting the dataset into two clusters using K-Means, followed by applying SMOTE separately within each cluster. By generating synthetic samples within more homogeneous subgroups, this method aims to create a representative, balanced dataset while reducing the risk of overfitting. Since SMOTE is applied within structurally similar data points, the synthetic instances better preserve the original data distribution, potentially leading to improved model generalization. The results after applying clustering and SMOTE are as follows (Table 7):

Table 7: Methodology 2: Results with clustering prior to SMOTE

Model	Accuracy	Precision	Specificity	Sensitivity	F1 Score	AUC score	RMSE	Log Loss	MCC	Cohen's Kappa	Processing time (sec)
LR	0.9288	0.9072	0.8944	0.9609	0.9333	0.9681	0.2668	0.2147	0.8588	0.8572	0.30
SVM	0.9694	0.9430	0.9379	1.0000	0.9707	0.9907	0.1750	0.0957	0.9404	0.9387	3.40
DT	0.9656	0.9518	0.9516	0.9800	0.9657	0.9658	0.1855	1.1882	0.9316	0.9312	0.56
RF	0.9769	0.9605	0.9561	0.9964	0.9781	0.9975	0.1520	0.1063	0.9544	0.9537	0.79
XGBoost	0.9920	0.9850	0.9851	0.9990	0.9920	0.9996	0.0895	0.0273	0.9841	0.9840	0.68

Applying clustering before SMOTE leads to noticeable improvements across all performance metrics compared to using SMOTE alone. Accuracy increases for all models, with XGBoost reaching 0.9920. Precision, specificity, and F1-score also improve, showing better classification balance. The AUC score rises, indicating stronger model discrimination, while RMSE and log loss decrease, reflecting more reliable predictions. Clustering allows SMOTE to generate synthetic samples that

better represent the data structure, leading to improved generalization and reduced overfitting. This method helps models learn more meaningful patterns, enhancing their robustness and predictive power compared to SMOTE without prior clustering.

4.4.4 Misclassification rate comparison.

This section compares misclassification rates across various models and methodologies, including the baseline (no balancing techniques), SMOTE Application, and the combination of clustering and SMOTE. The goal is to evaluate the effectiveness of these balancing techniques in reducing misclassification. The results, derived from confusion matrices, reveal the percentage of misclassified observations for each model and methodology (TABLE 8 and TABLE 9).

Table 8: Methodology 1: Confusion Matrix Analysis

Model	TN	FP	FN	TP
LR	932	124	62	1004
SVM	971	89	0	1062
DT	1009	71	25	1017
RF	970	67	10	1075
XGBoost	1029	45	0	1048

Table 9: Methodology 2: Confusion Matrix Analysis

Model	TN	FP	FN	TP
LR	915	108	43	1056
SVM	982	65	0	1075
DT	1022	52	21	1027
RF	980	45	4	1093
XGBoost	1056	16	1	1049

The clustered SMOTE Approach consistently reduces misclassification rates, with the most significant improvement seen in the XGBoost model, where the misclassification rate drops to as low as 0.80%, corresponding to just 17 misclassified observations out of a total of 2122 (Table 6., FIGURE 3.). This is a significant reduction compared to the 2.12% or 45 misclassified observations without clustering (SMOTE Application only), all while keeping the other parameters of the model unchanged. This highlights that combining clustering with SMOTE not only enhances model performance but also improves their ability to classify observations accurately, especially for models like XGBoost, which benefit most from this approach.

5. CONCLUSION

The study substantially assists in predicting financial distress since it solves the continuing challenge of unbalanced classes within financial risk assessment data. Traditional predictive models deal with poor financial distress identification because their capabilities are affected by the mismatch between

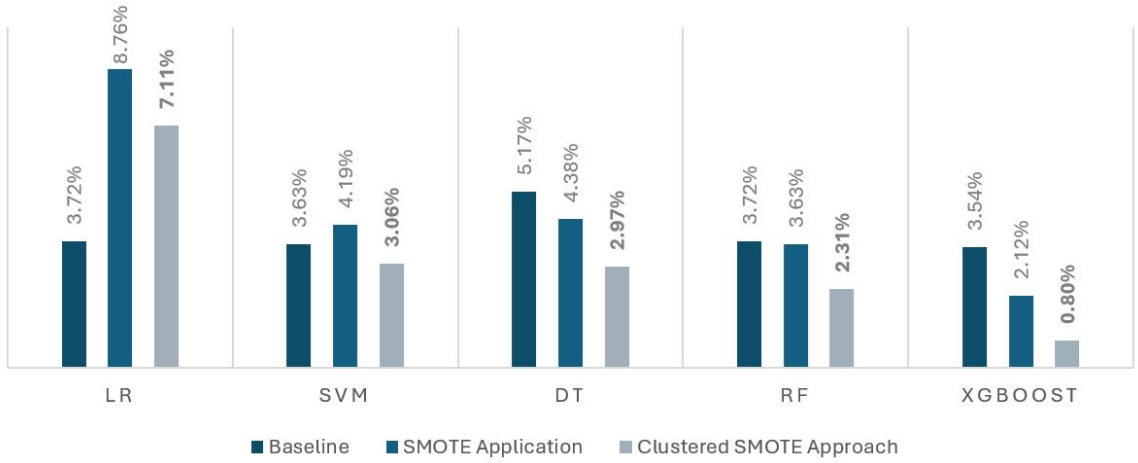

Figure 3: Misclassification Rates Comparison – Shows the overall percentage of misclassified observations, calculated as (FP + FN) divided by the total number of observations from the remaining entries in the confusion matrix.

thriving firms and struggling companies. Combining Synthetic Minority Over-sampling Technique (SMOTE) with K-means clustering introduces an advanced model that strengthens data balancing capabilities to enhance predictive accuracy and effectiveness decisively. This study establishes its main strength by combining K-means clustering with SMOTE application in novel ways. The proposed methodology creates more useful synthetically generated samples that concentrate on specific data subsets to improve expected financial distress predictions.

The specified data subset target allows this method to generate more meaningful synthetic samples, enhancing model accuracy for financial distress prediction. Implementing clustering together with SMOTE generates improved accuracy levels in financial distress predictions through more effective detection of distressed company characteristics while preserving comprehensive classification effectiveness.

Our methodology demonstrates exceptional improvements compared to previous research. The clustered SMOTE approach with XGBoost achieved an outstanding Accuracy of 99.20%, Precision of 98.50%, Sensitivity of 99.90%, and F1 score of 99.20%, alongside an AUC score of 99.96% and an RMSE of 0.0895. These results not only surpass those achieved in existing research, but also show enhanced specificity and deduced Log Loss. This recalls a significant advancement in predictive accuracy and computational efficiency in financial distress prediction models. The clustered SMOTE method reduces misclassification errors, showing that the model has high reliability. The approach establishes 17 wrong predictions among 2122 records, thus achieving superior results when applied to data sets with low event frequencies. Predictive models in financial risk management debt significant value from accurate distressed company detection since it directly affects the real-world risk evaluation and decision-making processes.

Multiple key metrics demonstrate the model's strong generalization abilities, including accuracy, Precision and sensitivity, F1 score, AUC, and MCC. The model exhibits an accurate calibration through its minimal Log Loss value, which stops model overfitting while guaranteeing the reliability of predictions.

The research generates important findings about financial distress prediction, which help researchers build superior predictive models for financial risk evaluation. The work establishes innovative class imbalance handling to create a new field standard that boosts financial decision systems and risk assessment capabilities. According to the authors, analyzing financial distress patterns through time might provide new ways to forecast business failures. The current study treats financial distress as a static classification problem, where models predict a company's financial condition based on its financial and non-financial characteristics at a specific point in time, which has been used in previous researches. However, by re-framing the problem as a multivariate time series classification task, new patterns could emerge, improving the predictive power of the models. Incorporating time series analysis, recurrent neural networks (RNNs), Long Short-Term Memory (LSTM) networks, or dynamic adaptive models would allow the models to learn from temporal changes in company characteristics and external factors, offering a more dynamic and accurate forecast. This expansion would deepen the understanding of financial distress evolution over time, ultimately enhancing prediction accuracy and providing timely insights into the financial health of companies.

The presented research establishes better financial distress prediction methods by establishing new standards for balancing classes to deliver enhanced model predictive capability.

References

[1] Sun J, Li H, Fujita H, Fu B, Ai W. Class-Imbalanced Dynamic Financial Distress Prediction Based on Adaboost-SVM Ensemble Combined With SMOTE and Time Weighting. Inf Fusion. 2020;54:128-144.

[2] Aljawazneh H, Mora AM, Garcia-Sanchez P, Castillo-Valdivieso PA. Comparing the Performance of Deep Learning Methods to Predict Companies Financial Failure. IEEE Access. 2021;9:97010-97038.

[3] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic Minority Over-Sampling Technique. J Artif Intell Res. 2002;16:321-357.

[4] Elreedy D, Atiya AF. A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance. Inf Sci. 2019;505:32-64.

[5] Khushi M, Shaukat K, Alam TM, Hameed IA, Uddin S, et al. A Comparative Performance Analysis of Data Resampling Methods on Imbalance Medical Data. IEEE Access. 2021;9:109960-109975.

[6] Yang Y, Khorshidi HA, Aickelin U. A Review on Over-Sampling Techniques in Classification of Multi-Class Imbalanced Datasets: Insights for Medical Problems. Front Digit Health. 2024;6:1430245.

[7] Rahman HA, Wah YB, He H, Bulgiba A. Comparisons of ADABOOST, KNN, SVM and Logistic Regression in Classification of Imbalanced Dataset. Soft computing in data science. In: First International Conference SCDS Putrajaya Malaysia proceedings. Singapore: Springer. 2015:54-64.

[8] Liang J, Bai L, Dang C, Cao F. The K-Means-Type Algorithms Versus Imbalanced Data Distributions. IEEE Trans Fuzzy Syst. 2012;20:728-745.

[9] Malinen MI, Fränti P. Balanced k-means for clustering. InStructural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshop. S+ SSPR 2014. Springer Berlin Heidelberg.2014:32-41.

[10] Rokach L, Maimon O. Clustering Methods. Data Mining and Knowledge Discovery Handbook. 2005:321-352.

[11] Pradipta GA, Wardoyo R, Musdholifah A, Sanjaya IN, Ismail M. SMOTE for Handling Imbalanced Data Problem: A Review. Sixth International Conference on Informatics and Computing (Icic). IEEE. 2021:1-8.

[12] Jadwal PK, Jain S, Pathak S, Agarwal B. Improved Resampling Algorithm Through a Modified Oversampling Approach Based on Spectral Clustering And SMOTE. Microsyst Technol. 2022;28:2669-2677.

[13] Le TT, Oktian YE, Kim H. XGBoost For Imbalanced Multiclass Classification-Based Industrial Internet of Things Intrusion Detection Systems. Sustainability. 2022;14:8707.

[14] Fan M, Mo Z, Zhao Q, Liang Z. Innovative Insights Into Knowledge-Driven Financial Distress Prediction: A Comprehensive XAI Approach. J Knowl Econ. 2024;15:12554-12595.

[15] Liu W, Fan H, Xia M, Pang C. Predicting and Interpreting Financial Distress Using a Weighted Boosted Tree-Based Tree. Eng Appl Artif Intell. 2022;116:105466.

[16] Hou GD, Tong DL, Liew SY, Choo PY. Exploring Random Forest Regression for Financial Distress Detection 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS). IEEE. 2024:7-12.

[17] Hua Z, Wang Y, Xu X, Zhang B, Liang L. Predicting Corporate Financial Distress Based on Integration of Support Vector Machine and Logistic Regression. Expert Syst Appl. 2007;33:434-440.

[18] Kim SY, Upneja A. Predicting Restaurant Financial Distress Using Decision Tree and AdaBoosted Decision Tree Models. Econ Modell. 2014;36:354-362.

[19] Jaffari AA, Ghafoor Z. Predicting Corporate Bankruptcy in Pakistan a Comparative Study of Multiple Discriminant Analysis (MDA) and Logistic Regression. Res J Fin Acc. 2017;8:81-100.

[20] Xie C, Luo C, Yu X. Financial distress prediction based on SVM and MDA methods: the case of Chinese listed companies. Qual Quant. 2011;45:671-686.

[21] Saha S, Sarker PS, Saud AA, Shatabda S, Hakim Newton MA. Cluster-Oriented Instance Selection for Classification Problems. Inf Sci. 2022;602:143-158.

[22] Tsai CF, Lin WC, Hu YH, Yao GT. Under-Sampling Class Imbalanced Datasets by Combining Clustering Analysis and Instance Selection. Inf Sci. 2019;477:47-54.

[23] Kuizinienė D, Krilavičius T. Under-Sampling Class Imbalanced Datasets by Combining Clustering Analysis and Instance Selection balancing Techniques for Advanced Financial Distress Detection Using Artificial Intelligence. Electronics. 2024;13:1596.

[24] Sreedharan, Khedr AM, El Bannany M. A Multi-Layer Perceptron Approach to Financial Distress Prediction With Genetic Algorithm. Autom Control Comput Sci. 2020;54:475-482.

[25] Sandag GA. A Prediction Model of Company Health Using Bagging Classifier. JITK Jurnal Ilmu Pengetahuan Dan Teknologi Komputer. 2020;6:41-46.

[26] Al Ali AI, S SR, Khedr AM. Enhancing Financial Distress Prediction Through Integrated Chinese Whisper Clustering and Federated Learning. J Open Innov Technol Mark Complexity. 2024;10:100344.

[27] Rahman MJ, Zhu H. Predicting Financial Distress Using Machine Learning Approaches: Evidence China. J Contemp Acc Econ. 2024;20:100403.

[28] Jan CL. Financial Information Asymmetry: Using Deep Learning Algorithms to Predict Financial Distress. Symmetry. 2021;13:443.

[29] https://www.kaggle.com/datasets/shebrahimi/financial-distress

[30] Younus ZS, Mohamad D, Saba T, Alkawaz MH, Rehman A, et al. Content-Based Image Retrieval Using Pso and K-Means Clustering Algorithm. Arab J Geosci. 2015;8:6211-6224.

[31] Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, et al. Clustering Algorithms: A Comparative Approach. PLOS One. 2019;14:e0210236.

[32] Shutaywi M, Kachouie NN. Silhouette Analysis for Performance Evaluation in Machine Learning With Applications to Clustering. Entropy. 2021;23:759.

[33] Fernández A, Garcia S, Herrera F, Chawla NV. SMOTE For Learning From Imbalanced Data: Progress and Challenges Marking the 15-Year Anniversary. J Artif Intell Res. 2018;61:863-905.

[34] Singh ND, Dhall A. Clustering and Learning From Imbalanced Data. 2018. ArXiv preprint https://arxiv.org/pdf/1811.00972

[35] Elreedy D, Atiya AF, Kamalov F. A Theoretical Distribution Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Imbalanced Learning. Mach Learn. 2024;113:4903-4923.

[36] Wang L, Han M, Li X, Zhang N, Cheng H. Review of Classification Methods on Unbalanced Data Sets. IEEE Access. 2021;9:64606-64628.

[37] He H, Garcia EA. Learning From Imbalanced Data. IEEE Trans Knowl Data Eng. 2009;21:1263-1284.

[38] Haq MI, Ramadhan FD, Az-Zahra F, Kurniawati L, Helen A. Classification of Water Potability Using Machine Learning Algorithms. International Conference on Artificial Intelligence and Big Data Analytics. IEEE. 2021.

[39] Nair JP, Vijaya MS. Predictive Models for River Water Quality Using Machine Learning and Big Data Techniques-a Survey. International Conference on Artificial Intelligence and Smart Systems (ICAIS). IEEE. 2021:1-5.

[40] Charbuty B, Abdulazeez A. Classification Based on Decision Tree Algorithm for Machine Learning. J Appl Sci Technol Trends. 2021;2:20-28.

[41] Wu J, Chen XY, Zhang H, Xiong LD, Lei H, et al. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. J Electron Sci Technol. 2019;17:26-40.

[42] Ali ZA, Abduljabbar ZH, Tahir HA, Sallow AB, Almufti SM. eXtreme Gradient Boosting Algorithm With Machine Learning: A Review. Acad J Nawroz Univ. 2023;12:320-334.

[43] Liang J. Confusion Matrix: Machine Learning. POGIL Activity Clearinghouse. 2022;3.