Predicting Thai Listed Company Financial Distress by Machine Learning and Synthetic Minority Oversampling Technique and Hybrid Resampling Techniques

Phatchara Plypichit [1,2,*] and Supranee Lisawadi [1]

[1] Department of Mathematics and Statistics, Thammasat University, Pathum Thani, Thailand
[2] Securities and Exchange Commission, Thailand
Email: phatchara.ply@dome.tu.ac.th (P.P.); supranee@mathstat.sci.tu.ac.th (S.L.)
*Corresponding author

Manuscript received November 21, 2023; revised January 15, 2024; accepted February 7, 2024; published August 5, 2024.

***Abstract*—Nowadays, the techniques of machine learning have been widely adopted internationally for corporate financial distress prediction. The problem of unbalanced class distribution in classifying financial distress of listed companies on the Stock Exchange of Thailand (SET) may be addressed by implementing effective instruments by oversampling techniques and a mix of resampling methods. Many research studies have used different methods and financial ratios to classify financial distressed companies and understand the negative impact of major indicators on company financial position. This research analyzes data from financial statements gathered from 650 publicly listed firms on the SET during 2022, identify distressed companies by three consecutive years of negative earnings. Resampling was done to improve the g-mean and balanced accuracy of the three classifiers (C5.0, PART, and Generalized Linear Model (GLM)). Borderline Synthetic Minority Oversampling Technique (BLSMOTE) combined with C5.0 produces the highest median g-mean and balanced accuracy scores of 70.27% and 73.79%, respectively. In addition, SMOTE combined with Enhanced Nearest Neighbor (ENN), One-Sided Selection (OSS), and Tomek links increase the C5.0 g-mean scores compared to standalone SMOTE. This comparative analysis emphasizes oversampling and hybrid resampling effectiveness in addressing class imbalance in financial distress prediction. These findings have practical implications for stakeholders and decision-makers, suggesting the use of machine learning models with resampling techniques for early financial distress detection.**

***Keywords*—imbalanced data, financial distress prediction, classification, resampling methods, machine learning**

I. Introduction

The economy in the 1st decade of the 21st century is greatly affected by financial crises involving publicly traded companies, such as Enron and Lehman Brothers. Events of bankruptcy disrupt the financial market and have an impact on connected industries within the nation. The quantity of insolvent businesses reflects the level of economic growth and stability in a nation. Financial distress is when a company cannot produce enough cash flow, sustain profitability, or fulfill its responsibilities, resulting in significant societal issues like economic downturn and increasing unemployment. The 2008 global financial crisis heightened financial institutions' emphasis on evaluating credit risk and predicting financial distress (Huang and Yen, 2019). Financial distress is a major concern for everyone involved in the intricate economic landscape, as it puts companies at risk and results in losses for stakeholders (Liu *et al.*, 2022).

Financial distress prediction is highly important when making lending choices and determining the success of fund providing of financial institutions. Essential financial reports, like the balance sheet, income statement, cash flow statement, and equity statement, give insights into a company's activities and financial state. This information, particularly cash flows, can be utilized to anticipate potential financial distress. Liu *et al.* (2022) employed various methods to classify financially difficult companies, such as considering consecutive years of financial difficulties or poor performances. Different periods of financial information have been employed in many research studies to predict financial distress, with time frames varying from one year before facing financial troubles to four years before being removed from listing. The research gathered financial ratios from different periods leading up to delisting, to understand how certain financial indicators negatively affect a company's financial condition.

Traditional statistical approaches for forecasting bankruptcy encounter some limitations and restrictions because of their rigid assumptions. Conventional credit risk assessment methods depend on the subjective evaluations of human experts, which means they are responsive rather than anticipatory. Multiple methods have been created to anticipate financial distress, including linear and logistic regressions, survival analysis, and multiple-criteria programming. However, these approaches often depend on assumptions that may not align with real-life scenarios (Huang and Yen, 2019). Traditional linear algorithms such as logistic regression and linear discriminant analysis fail to adequately address the non-linear characteristics of economic indicators. Machine learning models have demonstrated potential performance in converting financial ratios and creating reliable financial distress early warning systems (Liu *et al.*, 2022). Kordestani *et al.* (2011) and Nopphaisit and Likitwongkajon (2019) developed a predictive model that utilizes a company's cash flow statement to anticipate potential financial difficulties. The first research chose 70 struggling companies from the Tehran stock market, while the latter chose 95 distressed listed firms from the Stock Exchange of Thailand. Through the examination of the cash flow statement, the researchers discovered that the cash flow composition has the ability to anticipate financial distress.

Several researchers have suggested a model that utilizes common and uncomplicated machine learning models to anticipate forthcoming financial difficulties by analyzing

doi: 10.18178/ijtef.2024.15.3.776

financial ratios. Kristóf and Virág (2022) aim to enhance the existing body of knowledge by using specific machine learning techniques, namely C5.0 decision trees and neural networks, to forecast the probability of EU-27 banks encountering failure. The authors emphasize that earnings, capital sufficiency, and managerial competence are significant indicators of potential bank failure. Barboza *et al.* (2017) evaluated the effectiveness of machine learning models and conventional methods in forecasting financial distress. Support Vector Machines (SVMs), bagging, boosting, and Random Forests (RF) were discovered to have roughly 10% higher accuracy compared to conventional methods such as Logistic Regression (LR), discriminant analysis, and Neural Networks (NNs). The models used six financial variables, which included operating margin, changes in return on equity, changes in price-to-book ratio, and measures of growth such as assets, sales, and number of employees. RF was the most successful out of all the models. Huang and Yen (2019) examined six different machine learning techniques using real-world data from publicly traded companies in Taiwan. These methods included supervised models, an unsupervised classifier called Deep Belief Network (DBN), and a combination of DBN and Support Vector Machine (SVM). By utilizing financial indicators from the companies' statements, the research discovered that the XGBoost algorithm was the most accurate in predicting financial difficulties. The hybrid DBN-SVM model also showed superior forecast accuracy compared to using either model alone, suggesting that machine learning can effectively predict financial distress.

The class distribution of financial distress and general risk events frequently involves imbalances in data distribution (Kou *et al*., 2022). The issue of class imbalance and the enhancement of classifier performance have been effectively addressed through the utilization of resampling methods, as demonstrated by Batista *et al*. (2004) and He and Garcia (2009) extensively employed combinations of sampling methods and various classifier combinations to resolve class imbalance problems. Furthermore, research has concentrated on the hybrid resampling methods and classification algorithms to tackle class imbalance problems in credit assessments (Roy, 2018; Raghuwanshi, 2020; Koziarski, 2020; Kou, 2022). To illustrate, Sun *et al.* (2018) introduced a decision tree ensemble that combines the Synthetic Minority Over-sampling Technique (SMOTE) and a bagging algorithm in order to handle the problem of imbalanced enterprise credit classifications. This model was tested on six combination models with data gathered from Chinese firms. Almost imbalanced class handling models outperforms pure model. Kou *et al*. (2022) introduced an under-sampling method that combines distance measurement and majority class clustering offering a solution for imbalanced credit default problems.

Nowadays, modern machine learning methods have become prevalent in predicting corporate financial distress. These methods currently hold sway in the realm of corporate financial distress predictions. Recent investigations into financial distress prediction have yielded impressive results through the utilization of machine learning classifier methods, whereas logistic regression methods have retained their popularity and have generally been deemed reliable (Kristóf and Virág, 2022). A recent study has implemented financial distress prediction in the Chinese market, which is different from developed country markets, and proposed a model to determine which financial factors are more likely to affect companies and turn into financial distress during the COVID-19 crisis (Ding *et al*., 2023). Moreover, an imbalanced data problem is still needed to undertake preliminary processing of data before allowing classifiers to perform such as oversampling, under-sampling, or hybrid resampling. Hence, the objective of this study was to provide novel empirical models for effective corporate financial distress prediction. Multivariate classification models, including C5.0, PART, and GLM, were developed to anticipate the likelihood of distress among corporates listed in the Stock Exchange of Thailand. These models categorize corporates as either distressed or non-distressed firms in the following year based on financial factors such as earnings, all cash flow activities, and management activities, and then evaluate each model and compare its performance respectively. In particular, we tested whether the resampling-implemented model can significantly enhance the model without handling imbalance problems through a statistical analysis.

II. Methods

A. *Financial Distress Variables*

Financial statements are official documents that provide details of a company's financial transactions. Publicly traded companies must create four fundamental financial documents: balance sheet, income statement, statement of equity changes, and cash flow statement. These four fundamental financial statements offer details about the outcomes of the company's activities, its financial status, and cash movements. In this research, we organized financial ratio variables into four categories, which were suggested by several mentioned studies, and additionally placed specific emphasis on each cash flow activity variable. The all variables for companies in the listed domain are obtained from the SETSMART database. The selected variables with explanations are displayed in Table 1.

We gathered and examined the most up-to-date financial statement data for the 2022 period from 650 firms publicly listed in the Stock Exchange of Thailand. To identify distressed companies in the stock market, we apply a criterion that requires three consecutive years of negative earnings, as specified by Nopphaisit and Likitwongkajon (2019) and Jantadej (2006). Several researchers have studied and examined financial distressed companies including many industry groups, excluding those from the financials industry group and the Property Fund & REITs sector in the Property & Construction industry group. These selected companies have been listed on the Stock Exchange of Thailand for at least three consecutive years. Table 2 provides a summary of the industry category proportions, whereas Table 3 presents statistical data comparing distressed and non-distressed firms. The general process and research area are shown in Fig. 1.

Table 1. Selected independent financial variables for distressed firm prediction

Type	Variables	Financial ratio	Variables	Financial ratio
Liquidity	V_1	Current ratio	V_2	Quick ratio
Solvency	V_3	Debt to Equity	V_4	Interest Coverage
	V_5	Operating Cash Flow to Total Liabilities	V_6	Investing Cash Flow to Total Liabilities
	V_7	Financing Cash Flow to Total Liabilities	V_8	Net Cash Flow to Total Liabilities
Operational	V_9	Fixed Asset	V_{10}	Total Asset Turnover
Capabilities	V_{11}	Inventory Turnover	V_{12}	Average Sale Period
	V_{13}	Accounts Receivable Turnover	V_{14}	Average Collection Period
	V_{15}	Account Payable Turnover	V_{16}	Average Payment Period
V_{17}		Cash Cycle		
Profitability	V_{18}	Return on Asset	V_{19}	Return on Equity
	V_{20}	Gross Profit Margin	V_{21}	EBIT Margin
	V_{22}	Net Profit Margin	V_{23}	Operating Cash Flow to Total Asset
	V_{24}	Investing Cash Flow to Total Asset	V_{25}	Financing Cash Flow to Total Asset
	V_{26}	Net Cash Flow to Total Asset		

Table 2. Summary of the industry categories of the distressed and non-distressed companies

Industry categories	Non-distressed firms	%	Distressed firms	%
Agro & Food Industry	59	9.74%	6	13.64%
Consumer Products	45	7.43%	4	9.09%
Industrials	122	20.13%	6	13.64%
Property & Construction	114	18.81%	16	36.36%
Resources	67	11.06%	3	6.82%
Services	153	25.25%	8	18.18%
Technology	46	7.59%	1	2.27%
Total	606	100.00%	44	100.00%

Table 3. Statistical summary of financial ratio variables

	Non-distressed firms		**Distressed firms**		**Mann-Whitney U Test**
	Median	S.D.	Median	S.D.	Z-value
V1	1.7500	8.6908	1.0350	7.2034	6.0888***
V2	0.8400	6.7492	0.3850	2.2260	6.1399***
V3	0.7100	3.2034	1.3600	3.2138	1.9049
V4	5.9450	139139.7410	−0.1500	160.9166	6.5115***
V5	0.1295	1.0533	0.0014	0.4002	5.0665***
V6	-0.0700	4.3159	−0.0164	1.5275	3.4125***
V7	-0.0704	1.0139	−0.0055	0.5689	5.6284***
V8	0.0067	5.1258	0.0020	0.9348	3.7251***
V9	2.4500	18.3508	1.6200	9.2780	3.656***
V10	0.6900	0.7162	0.3300	0.3779	2.3609***
V11	5.8800	441.7204	5.2750	30.3435	4.1154***
V12	62.0650	822.9729	69.4100	2333.3117	1.2385
V13	6.9350	172.0737	5.1200	45.0680	1.2371
V14	52.6200	4688.4126	71.2800	3075.8892	1.9049
V15	5.7800	22.3328	3.5550	3.7779	1.9036
V16	63.1950	814.0520	102.7000	3494.3691	−4.519***
V17	54.9100	4110.1552	45.6300	1689.9865	−4.52***
V18	5.2450	9.7329	−0.8750	12.0934	−0.1786
V19	6.5900	33.6199	−6.2400	91.8242	−4.052***
V20	21.7800	18.0781	16.6550	20.2327	−2.158***
V21	7.9550	710.1763	−2.1600	67.4564	−2.227***
V22	5.0550	720.9448	−12.6550	85.8049	−0.1696
V23	0.0572	0.0972	0.0021	0.1183	−4.541***
V24	−0.0292	0.0921	−0.0115	0.1601	−2.249***
V25	−0.0261	0.1109	−0.0065	0.1350	−2.902***
V26	0.0034	0.0803	0.0015	0.0983	−0.531

Notes: *** p-value < 0.05 and means the medians of the two classes are significantly different at 0.05 level.

In this research, we would like to utilize more simple, well-known, and effective machine learning models to compare prediction accuracy and classify targeted distressed firms. As mentioned in the previous section, many papers have used bagging methods such as decision tree (C5.0), random forest, or rule-based methods such as PART to classify and predict binary classification problems and compare those machine learning models with traditional logistic regression model (Barboza, 2017; Sun, 2018; Kristóf, 2022). The resampling techniques and the hybrid resampling methods are currently interesting in the field of prediction science to tackle class imbalance problems in credit evaluation in several papers (Roy, 2018; Raghuwanshi, 2019, 2020; Koziarski, 2020; Kou, 2022). Therefore, we would participate from the literature that which model with or without the resampling method is the most efficient in accurately predicting financial distress firms.

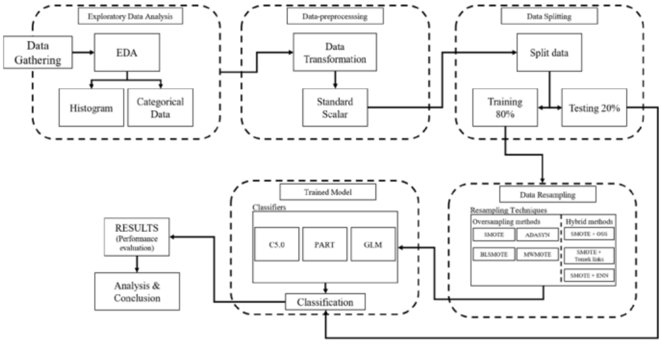

Fig. 1. Research workflow.

B. *Resampling Techniques*

1) *SMOTE*

The primary technique employed for oversampling in imbalanced datasets is Synthetic Minority Over-sampling Technique (SMOTE) which generates additional instances of the minority class by creating synthetic examples, instead of simply duplicating existing samples. This method increases the area in which the minority class can be classified as part of its decision region. This algorithm is employed to create artificial samples, working within the "feature space" instead of the "data space". The parameter allows adjustment of the extent of over-sampling. To produce synthetic samples, the process involves calculating the dissimilarity between the feature vector being examined and its closest neighbor. This calculated difference is then multiplied by a random value ranging from 0 to 1 and subsequently added to the original feature vector (Chawla *et al.*, 2002).

This method compels the decision area of the smaller class to become broader, leading to larger and less detailed decision regions. the SMOTE is used to create synthetic data points s, according to:

$$s = x + \alpha \times (y - x)$$

where α is a random number in the range [0,1].

2) *ADASYN*

Adaptive Synthetic Sampling method (ADASYN) is designed to tackle the issue of imbalanced data sets in machine learning. It offers a solution to decrease the bias caused by the imbalance in class distribution and dynamically adjusts the classification decision boundary to focus on difficult examples. This process creates additional synthetic data for the minority class with a focus on examples that are more difficult to learn by implementing a weighted distribution for these hard examples that aims to enhance the learning performance. This can be advantageous in real-life situations where there is a high presence of imbalanced data sets, such as in detecting fraud, diagnosing medical conditions, and identifying anomalies. Simulation analyses have shown that ADASYN is an effective method on various machine learning datasets and can be an effective approach for addressing imbalanced learning challenges, offering a potential remedy for researchers and professionals working in the working domain of data mining, artificial intelligence, and machine learning (Han *et al.*, 2005).

3) *BDLSMOTE*

Borderline-SMOTE (BDLSMOTE) is a data augmentation technique used in machine learning. Unbalanced data sets can be found in many areas of data mining, and it is vital to solve the imbalance issue to ensure precise analysis and prediction. Borderline-SMOTE is an over-sampling method that is highly valued for its effectiveness in specifically increasing the number of minority samples that are close to the borderline. The experimental findings demonstrate that the newly introduced methods outperform traditional techniques such as SMOTE and random over-sampling in terms of True Positive (TP) rate and F-value which is one of the popular evaluation metrics for handling imbalance problems that works in the combination of recall and precision metrics and incorporates the true positive rate accuracy from the minority class. This method is designed to identify the boundary examples of the smaller group and produce new artificial examples using these boundary examples. It should be emphasized that this method only increases the number or intensity of the boundary and nearby points of the smaller group, which sets it apart from SMOTE (He *et al.*, 2008).

4) *MWMOTE*

Modified Weighted Minority Oversampling Technique (MWMOTE) is a technique used to balance imbalanced datasets by creating artificial samples for the minority class. MWMOTE is a technique used to find and assign weights to minority class samples that are difficult to learn, based on their distance from the nearest majority class samples. Afterwards, it creates synthetic instances using a clustering method on the informative minority class samples, ensuring all generated samples lie inside minority class cluster (Barua *et al.*, 2014).

The combination of under-sampling and SMOTE can be a method utilized to address imbalanced datasets. This process includes randomly increasing the number of instances in the minority class and eliminating instances from the majority class to reduce the noise sample and borderline samples (Pristyanto *et al.*, 2017). Thus, we suggest three combination models in this study utilizing SMOTE and under-sampling approaches as following:

5) *SMOTE+OSS*

The One-Sided Selection (OSS) method is a strategy employed to address the issue of imbalanced class distribution by minimizing the possibility of losing crucial information present in the dataset. This principle allows dividing the majority class into four categories: noise, borderline, redundant, and safety samples. The OSS algorithm will eliminate a sample from the majority class that is surrounded by the minority class which is referred to noise samples. This algorithm will also delete a borderline sample which is situated in the middle between the two classes (Pristyanto *et al.*, 2017).

6) *SMOTE + ENN*

The combination of SMOTE and the Wilson's Enhanced Nearest Neighbor Rule (ENN) can handle the issue of class imbalance in machine learning training data. This technique merges the SMOTE oversampling technique with the ENN data cleaning technique. The approach operates by identifying samples that are wrongly classified by their three closest neighbors and eliminating them from the training dataset. The method seeks to balance the training data's quality and minimize the influence of class imbalance on the learning system by eliminating these misclassified examples (Ali and Smith, 2006).

7) *SMOTE+ TOMEK*

Tomek links serve as a technique for handling unbalanced class problems in machine learning. They can be utilized either as a means of under-sampling or for the purpose of data cleaning. On the other hand, only instances that belong to the majority class are removed. The learning algorithm will be compelled to give more attention to the minority category, which is frequently the category of concern. However, we utilize it as a technique to clean our data by removing examples from both classes. This approach cleans the over-sampled training set produced by SMOTE by employing Tomek links. Tomek links refer to pairs of examples from different categories that are the nearest to one another, and their removal can enhance the classes distinction. In order to create a balanced data set with distinct class clusters, not only majority class examples are eliminated but the examples from both classes are also removed (Ali and Smith, 2006).

C. *Machine Learning Models*

1) *C5.0*

Decision trees are a widely used methodology for data mining that offers both classification and predictive functions. Prominent algorithms for decision tree include ID3, C4.5, and C5.0, which progressively enhance the splitting rules, calculation methods, and rule generation. The C4.5 algorithm improves upon ID3 by utilizing a gain-ratio index to segment attributes, thereby mitigating the problem of excessive sub-trees. On the other hand, the C5.0 algorithm, being a commercial version of C4.5, further enhances rule generation, effectively handles large datasets, and exhibits superior speed and memory efficiency owing to the implementation of the Boosting method (Chen, 2011).

2) *PART*

The PART method is a classification algorithm that possesses an advantage over other methods in that it generates rules without necessitating global optimization. The partial decision tree derived from PART is a result of combining C4.5 and Repeated Incremental Pruning to Produce an Error Reduction (RIPPER) algorithm. This algorithm generates decision lists that are employed as a set of rules. As new data is introduced, it is compared to the existing rules, and if no matching rule is found, the corresponding clause is transferred. In the PART method, the dataset is first divided into a partial tree, and subsequently, tests are selected and divided into subsets. The development of these subsets is based on the average entropy. This process continues until a subset expands and reaches a leaf, which other subsets that remain unexpanded are selected in the subsequent steps. The optimal leaf is then identified as a rule. Thus, the PART algorithm produces a set of rules that can be utilized to classify testing data based on the patterns supervised from the training data (Samadianfard, 2022).

3) *GLM*

The logit methodology is a statistical technique employed in failure prediction studies. Logistic regression, a statistical method used to predict a dichotomous dependent variable and is widely used in various fields such as finance, healthcare, and social sciences. It predicts the likelihood of an event occurring based on independent variables. The logistic regression equation is derived using the maximum-likelihood estimation to ascertain the significance of variables. Logistic regression models typically have a categorical dependent variable with two or more levels. In this study, the dependent variable is whether a company is facing financial distress by labeling the dependent variable that the firm is encountering three consecutive years of negative earnings as 1, and several financial ratios are employed as independent variables to predict financial distress. A two-class system is coded using a 0/1 response, where $y_i = 1$ indicates that the firm is encountering financial distress in the next 1 year for the first class and $y_i = 0$ indicates that the firm is in healthy position for the second class. The relationship between p and x is expressed through the logit transformation, given by the Eq. (1):

$$Logit(p) = \ln\left(\frac{p}{1-p}\right) = f(x,\beta) = \beta^T x \cdot \qquad (1)$$

The logit transformation is also referred to as the log-odds transformation. The vector of coefficients of the model is denoted by $\beta = (\beta_i, \beta_i, \beta_i, ..., \beta_i)$, and β^T is the transpose vector. The expression p/(1-p) is referred to as the odds-ratio. The unknown regression coefficients β_i, which must be estimated from the data, are directly interpretable as log-odds ratios or, in terms of $\exp(\beta_i)$, as odds ratios. That log-likelihood for n^T observations is (Chen, 2011),

$$l(\beta) = \sum_{l=1}^{nT} \left\{ yl\beta^T xl + \log(1+e^{\beta^T xl}) \right\} \cdot \qquad (2)$$

Table 4. Confusion matrix for binary prediction the distressed and non-distressed companies

	Positive prediction	**Negative prediction**
Positive class	True Positive (TP)	False Negative (FN)
Negative class	False Positive (FP)	True Negative (TN)

Table 5. Classifier evaluation methodologies used in this study

Metric	Equation
Overall Accuracy	$\frac{TN+TP}{(TN+TP+FN+FP)}$
Specificity	$\frac{TN}{(TN+FP)}$
Sensitivity	$\frac{TP}{(TP+FN)}$
G-mean	$\sqrt{sensitivity \times specificity}$
Balanced Accuracy	$\frac{sensitivity+specificity}{2}$

D. *Performance Metrics*

In this particular investigation, we assess the classification performance based on an analysis of the confusion matrix. The confusion matrix is a tabular representation that shows the number of true positives, true negatives, false positives, and false negatives for a binary class. An example of a confusion matrix for a two-class problem, consisting of positive and negative class values can be found in Table 4 By examining the confusion matrix, one can extract various commonly used metrics for evaluating the performance of learning systems. In the context of binary classification, the output is either positive or negative. One such metric is Accuracy, which is calculated by dividing the sum of true positives and true negatives by the sum of true positives, false negatives, false positives, and true negatives. These metrics play a crucial role in the evaluation of learning systems, as they offer a quantitative measure of the system's ability to correctly classify instances into their respective classes. Table 5 provides an overview of the performance evaluation criteria for each method studied.

Many researchers argue that the overall accuracy metric may not be effective when dealing with imbalanced datasets (Sokolova, 2006; Gu, 2009; Akosa, 2017; Chicco, 2020). Thus, in this investigation, we employ g-mean and balanced accuracy as main metrics that address the prediction problems of the minority class in the testing dataset. By comparing the performance of different learning systems using these metrics, researchers can ascertain which system is most effective for a given purpose.

III. Experimental Results

A. *Class Balancing Process*

To address the issue of imbalanced datasets, the researchers employ various methods of dataset balancing, including oversampling and combinations of oversampling and under-sampling techniques. Specifically, the researchers utilize the R programming language, an open-source tool widely employed by statisticians, for implementing these resampling techniques. Fig. 2 provides a visualization of the lass distribution in the original dataset, highlighting the imbalanced nature of the problem that the majority class samples dominate the minority samples, indicating the imbalance. However, after applying the hybrid method for class balancing, the class distribution is adjusted, leading to an increase in the values of g-mean and balanced accuracy when employing different classification algorithms. It is vital to achieve a balanced dataset, as an imbalanced dataset can significantly degrade the performance of various data classification algorithms.

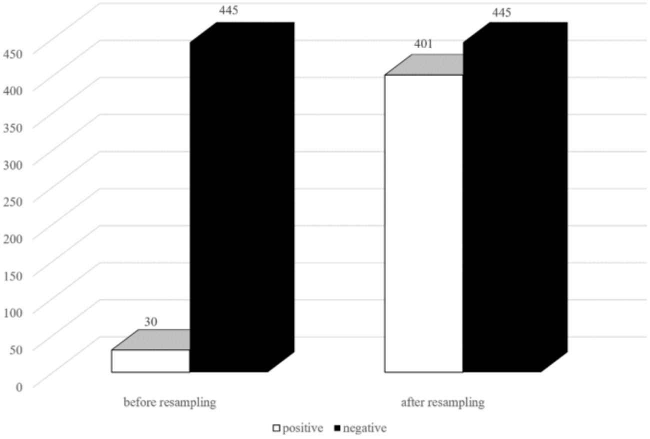

Fig. 2. Illustration of before and after class balancing process.

The oversampling techniques employed in this study include SMOTE, ADASYN, BLSMOTE, and MWMOTE, which generate synthetic samples for the minority class using specific algorithms based on the existing samples. In addition to oversampling, the researchers also utilize under-sampling techniques in combination with oversampling, such as SMOTE with OSS, Tomek, ENN, to remove some samples from the majority class and balance the class distribution. By employing this combination of methods, the study aims to address the issue of unbalanced class distribution in data mining processes. It is important to note that oversampling can potentially lead to overfitting, while under-sampling may result in the loss of important data.

B. *Performance Evaluation*

After finishing the task of handling imbalanced data, the next step involves the classification and performance evaluation process. This research utilized three commonly used algorithms (C5.0, PART, and GLM) to forecast the financial distressed companies using the provided dataset by training the data and then predict with testing data. The dataset was split into two parts, one for training the model and the other for testing it. 80% of the complete dataset was used for training, and the remaining 20% was set aside for testing. The models were initially assigned default parameters, and two separate 5-fold cross-validations were used as the resampling during the training process. After repeating 30 iterations, the research then evaluates the mean g-mean, balanced accuracy and accuracy of the testing set from three mentioned algorithms, both with and without class balancing procedure. The comparable findings are displayed in Tables 6 and 7. It has been determined that incorporating resampling methods such as oversampling or combination techniques can improve the average g-mean and balanced accuracy values of the three algorithms.

Additionally, the accuracy value of all techniques almost drops after implementing resampling technique. However, it is pointed out that the significant importance of the g-mean, the balanced accuracy and the accuracy for each algorithm can be determined through statistical testing.

C. *Comparative Analysis*

The significant difference in method comparison can be determined by statistical tests and hypothesis testing to confirm the significantly improved performance between the algorithms compared. A t-test was used to determine whether the metric value after implementing the resampling method (μ_2) differs from the value without handling imbalanced data (μ_1) with a significance level of 0.05. The null and alternative hypotheses to test the level of significance are as follows:

$$H_0: \mu_2 = \mu_1$$

$$H_1: \mu_2 > \mu_1$$

The average g-mean and balanced accuracy were tested to determine whether they significantly differed from the benchmark metrics. Table 8 shows the metrics and t-calculated value from hypothesis testing. Based on the statistical t-test, the addition of oversampling and combination methods can significantly improve the g-mean and balance accuracy scores of the C5.0 and PART algorithms to handle imbalanced data.

The alternative to parametric t-tests is the median test, which is a nonparametric test that determines whether the medians of the metrics of two or more independent samples are equal. In this study, we compared the metrics obtained using additional techniques with those obtained using the original testing data. Based on the test statistic and p-value, the median g-mean and balanced accuracy of the C5.0 algorithms with all techniques are significantly different from the median metrics evaluated using imbalanced testing data as illustrated in Table 8.

Table 6. Comparing mean and median G-Mean, Balanced accuracy and Accuracy metrics without handling imbalanced data

Algorithm	Mean G-Mean (%)	Median G-Mean (%)	Mean Balanced accuracy (%)	Median Balanced accuracy (%)	Mean Accuracy (%)	Median Accuracy (%)
C5.0	8.19%	0.00%	50.82%	50.00%	**90.97%**	**93.80%**
PART	24.37%	31.00%	52.62%	50.00%	81.56%	91.47%
GLM	14.98%	0.00%	49.33%	49.38%	72.24%	87.60%

Table 7. Comparing mean and median G-Mean, Balanced accuracy and Accuracy metrics with handling imbalanced data

Algorithm	Mean G-Mean (%)	Median G-Mean (%)	Mean Balanced accuracy (%)	Median Balanced accuracy (%)	Mean Accuracy (%)	Median Accuracy (%)
C5.0 (SMOTE)	48.17%	48.32%	62.58%	59.19%	83.26%	91.47%
PART (SMOTE)	50.46%	58.21%	62.93%	58.88%	73.84%	75.19%
GLM (SMOTE)	18.37%	12.86%	50.72%	50.41%	25.94%	7.75%
C5.0 (ADASYN)	49.23%	51.99%	62.31%	61.26%	82.80%	90.31%
PART (ADASYN)	44.28%	46.35%	59.00%	59.29%	66.11%	78.29%
GLM (ADASYN)	16.11%	9.09%	50.04%	50.00%	22.48%	6.98%
C5.0 (BLSMOTE)	**69.75%**	**70.27%**	**72.92%**	**73.79%**	78.71%	82.95%
PART (BLSMOTE)	50.67%	57.79%	59.69%	59.50%	61.92%	65.50%
GLM(BLSMOTE)	12.16%	0.00%	51.13%	50.00%	25.27%	7.75%
C5.0 (MWMOTE)	57.28%	64.33%	65.80%	68.33%	80.55%	85.27%
PART (MWMOTE)	58.88%	66.80%	65.66%	69.32%	62.65%	58.91%
GLM (MWMOTE)	14.26%	9.09%	50.29%	50.41%	28.06%	7.75%
C5.0 (SMOTE+OSS)	49.21%	50.36%	58.71%	58.55%	67.74%	73.26%
PART (SMOTE+OSS)	38.77%	34.62%	53.13%	52.53%	64.43%	72.09%
GLM (SMOTE+OSS)	22.72%	16.38%	51.34%	50.83%	55.94%	83.72%
C5.0 (SMOTE+ENN)	49.24%	50.36%	58.62%	58.55%	68.00%	73.26%
PART (SMOTE+ENN)	39.63%	39.24%	53.51%	52.53%	63.89%	70.40%
GLM (SMOTE+ENN)	22.72%	16.38%	51.34%	50.83%	55.94%	83.72%
C5.0 (SMOTE+Tomek)	53.07%	56.02%	61.18%	58.94%	74.57%	85.27%
PART (SMOTE+Tomek)	31.39%	34.02%	51.11%	50.00%	70.22%	77.52%
GLM (SMOTE+Tomek)	19.11%	12.86%	51.55%	50.83%	50.87%	48.99%

Table 8. Comparing test statistics of G-Mean and Balanced accuracy from hypothesis testing

Algorithm	Mean G-Mean (%)	T-test	Median G-Mean (%)	Test Statistic	Mean Balanced accuracy (%)	T-test	Median Balanced accuracy (%)	Test Statistic
C5.0 (SMOTE)	48.17%	7.6687***	48.32%	35.6229***	62.58%	6.7223***	59.19%	32.2667***
PART (SMOTE)	50.46%	3.9499***	58.21%	6.6667***	62.93%	4.3542***	58.88%	6.6667***
GLM (SMOTE)	18.37%	0.6814	12.86%		50.72%	1.2659	50.41%	
C5.0 (ADASYN)	49.23%	7.8672***	51.99%	32.8507***	62.31%	7.4623***	61.26%	29.4327***
PART (ADASYN)	44.28%	3.4878***	46.35%	4.2667***	59.00%	3.8686***	59.29%	4.2667***
GLM (ADASYN)	16.11%	0.2283	9.09%		50.04%	0.6082	50.00%	
C5.0 (BLSMOTE)	**69.75%**	16.3094***	**70.27%**	45.0667***	**72.92%**	15.5166***	**73.79%**	60.0000***
PART (BLSMOTE)	50.67%	4.7277***	57.79%	4.2667***	59.69%	3.0861***	59.50%	4.2667***
GLM(BLSMOTE)	12.16%	−0.6032	0.00%		51.13%	1.7120***	50.00%	
C5.0 (MWMOTE)	57.28%	9.5028***	64.33%	35.6229***	65.80%	8.5883***	68.33%	32.2667***
PART (MWMOTE)	58.88%	5.9411***	66.80%	26.6667***	65.66%	6.4523***	69.32%	21.6000***
GLM (MWMOTE)	14.26%	−0.1527	9.09%		50.29%	0.7930	50.41%	
C5.0 (SMOTE+OSS)	49.21%	9.3995***	50.36%	21.6000***	58.71%	4.5887***	58.55%	19.2881***
PART (SMOTE+OSS)	38.77%	2.5711***	34.62%	1.6685	53.13%	0.3161	52.53%	0.6007
GLM (SMOTE+OSS)	22.72%	1.5642	16.38%		51.34%	1.4571***	50.83%	
C5.0 (SMOTE+ENN)	49.24%	9.3970***	50.36%	21.6000***	58.62%	4.5949***	58.55%	19.2881***
PART (SMOTE+ENN)	39.63%	2.7033***	39.24%	1.6685	53.51%	0.5520	52.53%	1.0667
GLM (SMOTE+ENN)	22.72%	1.5642	16.38%		51.34%	1.4571***	50.83%	
C5.0 (SMOTE+Tomek)	53.07%	10.0904***	56.02%	19.2881***	61.18%	6.5001***	58.94%	19.2881***
PART (SMOTE+Tomek)	31.39%	1.2780	34.02%	1.0714	51.11%	−1.0470	50.00%	0.0000
GLM (SMOTE+Tomek)	19.11%	0.8827	12.86%		51.55%	1.9745***	50.83%	

Notes: *** p-value < 0.05 and means the mean or medians of the model are significantly different from benchmark at 0.05 level.

IV. CONCLUSIONS

The problem of unbalanced class distribution when classify the financial distressed companies listed on the Stock Exchange of Thailand has been successfully tackled by employing oversampling techniques and a mix of resampling methods. The utilization of the resampling approach in this study greatly significant enhances the g-mean and the balanced accuracy values of the classification algorithms (C5.0, PART, excluding GLM). The findings indicate that BLSMOTE with C5.0 produces the highest median g-mean and balanced accuracy scores, which are 70.27% and 73.79% respectively. Additionally, the combination of SMOTE with ENN, OSS, Tomek techniques lead to a rise in the C5.0's g-mean scores rising from 48.17% of standalone SMOTE to 49.21%, 49.24% and 53.07%, respectively. Similarly, the g-mean and the balanced accuracy values from all resampling techniques trained by C5.0 algorithm are significantly greater than values without handling imbalanced data. Moreover, this research proposes that the resampling technique should be used in conjunction with several machine-learning models in order to identify which is the most effective classifier for each specific combination.

The study provides results for the utilization of the BLSMOTE and C5.0 method over other resampling techniques due to its superior capability in predicting financial distress compared to the other two supervised methodologies. Additionally, the study reveals that the conventional logistic method cannot perform relatively well because this statistical technique is sensitive to multicollinearity, in which one independent variable in a multiple regression model can be perfectly correlated with another predictor. According to the GLM's g-mean and balance accuracy scores, it can be suggested that non restricted assumption model such as the more complex decision tree method can serve as an alternative predictor. Therefore, listed companies' financial statement data can be used to predict financial distress within a company by integrating various novel resampling techniques, machine learning-based methodologies to make the predictions.

In international application problems, the proposed models could be utilized as an alternative tool for class imbalance problems with testing performance with generally and commonly available financial ratios datasets. The results indicate the potential applicability and generalizability of the findings could also be applied to handle imbalance learning of financial distress problems in other contexts.

The comparative analysis conducted in this study highlights the significance of the improvements achieved by the oversampling and hybrid resampling approach, underscoring its effectiveness in addressing class imbalance in financial distress prediction. Furthermore, it provides practical implications for stakeholders and decision-makers, recommending the application of machine learning models with resampling techniques for early detection of financial distress.

Some considerable issues can be further addressed. First, further study is needed to expand the independent variables more than financial ratios, such as earnings management variables, auditor opinions, financial covenants, etc. Second, due to the limitation of datasets to test only one fiscal year, further study may test with many periods or between two or more events to evaluate the robustness of the models. Third, researchers can use and compare other imbalanced handling methods, such as cost-sensitive learning. Finally, as not implemented in this study, we may use feature selection methods to reduce redundant variables and processing time.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Phatchara Plypichit conducted the research, analyzed the data and wrote the paper; Supranee Lisawadi improved and edited the paper; both authors had approved the final version.

REFERENCES

Akosa, J.S. (2017). Predictive Accuracy: A Misleading Performance Measure for Highly Imbalanced Data. ***SAS Global Forum 942***, 2–5.

Ali, S., & Smith, K. A. 2006. On learning algorithm selection for classification. ***Applied Soft Computing Journal***, 6(2): 119–138.

Barboza, F., Kimura, H., & Altman, E. 2017. Machine learning models and bankruptcy prediction. ***Expert Systems with Applications***, 83: 405–417.

Barua, S., Islam, M. M., Yao, X., & Murase, K. 2014. MWMOTE - Majority weighted minority oversampling technique for imbalanced data set learning. ***IEEE Transactions on Knowledge and Data Engineering***, 26(2): 405–425.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. 2004. A study of the behavior of several methods for balancing machine learning training data. ***ACM SIGKDD Explorations Newsletter***, 6(1): 20–29.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. ***Journal of Artificial Intelligence Research***, 16(1); 321–357.

Chen, M. Y. 2011. Predicting corporate financial distress based on integration of decision tree classification and logistic regression. ***Expert Systems With Applications***, 38(9): 11261–11272.

Chicco, D., & Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. ***BMC Genomics***, 21(6).

Ding, S., Cui, T., Bellotti, A. G., Abedin, M. Z., & Lucey, B. (2023). The role of feature importance in predicting corporate financial distress in pre and post COVID periods: Evidence from China. ***International Review of Financial Analysis***, 90.

Gu, Q., Zhu, L., & Cai, Z. 2009. Evaluation Measures of the Classification Performance of Imbalanced Data Sets, In: Cai, Z., Li, Z., Kang, Z., Liu, Y. (eds) ***Computational Intelligence and Intelligent Systems. ISICA 2009. Communications in Computer and Information Science, Berlin, Heidelberg: Springer***, 51: 461–471.

Han, H., Wang, W., & Mao, B. 2005. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. ***Lecture Notes in Computer Science***, 3644(I): 878–887.

He, H., Bai, Y., Garcia, E. A., & Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. ***2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong***, 1322–1328.

He, H., & Garcia, E. A. 2009. Learning from imbalanced data. ***IEEE Transactions on Knowledge and Data Engineering***, 21(9): 1263–1284.

Huang, Y. P., & Yen, M. F. 2019. A new perspective of performance comparison among machine learning algorithms for financial distress prediction. ***Applied Soft Computing***, 83.

Liu, W., Fan, H., Xia, M., & Pang, C. 2022. Predicting and interpreting financial distress using a weighted boosted tree-based tree. ***Engineering Applications of Artificial Intelligence***, 116.

Jantadej, P. 2006. ***Using the combinations of cash flow components to predict financial distress***. Ph.D. dissertation, ETD collection for University of Nebraska - Lincoln.

Kordestani, G., Bakhtiari, M., & Biglari, V. 2011. Ability of combinations of cash flow components to predict financial distress. ***Business: Theory and Practice***, 12(3): 277-285.

Kou, G., Chen, H., & Hefni, M. A. 2022. Improved hybrid resampling and ensemble model for imbalance learning and credit evaluation. ***Journal of Management Science and Engineering***, 7(4): 511–529.

Koziarski, M., Woźniak, M., & Krawczyk, B. 2020. Combined Cleaning and Resampling algorithm for multi-class imbalanced data with label noise. ***Knowledge-Based Systems***, 204.

Kristóf, T., & Virág, M. 2022. EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks. ***Research in International Business and Finance***, 61.

Nopphaisit, N. & Likitwongkajon, N. 2019. Financial Distress Prediction through the Combination of Cash Flow Components of Listed Companies in the Stock Exchange of Thailand. ***NIDA Business Journal***, 25: 25–50.

Pristyanto, Y., Setiawan, N. A., & Ardiyanto, I. 2017. Hybrid resampling to handle imbalanced class on classification of student performance in classroom. ***2017 1st International Conference on Informatics and Computational Sciences (ICICoS), Semarang, Indonesia***, 207-212.

Raghuwanshi, B. S., & Shukla, S. 2019. Class imbalance learning using UnderBagging based kernelized extreme learning machine. ***Neurocomputing***, 329: 172–187.

Raghuwanshi, B. S., & Shukla, S. 2020. SMOTE based class-specific extreme learning machine for imbalanced learning. ***Knowledge-Based Systems***, 187.

Roy, A., Cruz, R. M. O., Sabourin, R., & Cavalcanti, G. D. C. 2018. A study on combining dynamic selection and data preprocessing for imbalance learning. ***Neurocomputing***, 286: 179–192.

Samadianfard, S. 2022. Evaluation of classification and decision trees in predicting daily precipitation occurrences. ***Water Supply***, 22(4).

Sokolova, M., Japkowicz, N., & Szpakowicz, S. 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation, In: Sattar, A., Kang, Bh. (eds) ***AI 2006: Advances in Artificial Intelligence. AI 2006, Berlin, Heidelberg: Springer***, 4304: 1015–1021.

Sun, J., Lang, J., Fujita, H., & Li, H. 2018. Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. ***Information Sciences***, 425: 76–91.