Accepted Paper

Uncorrected Proof

Advances in Mathematical Finance and Applications

Title: Developing Financial Distress Prediction Models Based on Imbalanced Dataset: Random Undersampling and Clustering Based Undersampling Approaches

Authors: Seyed Behrooz Razavi Ghomi, Alireza Mehrazin, Mohammadreza Shoorvarzi, Abolghasem Massihabadi

DOI: 10.22034/AMFA.2022.1956898.1743

Article ID: AMFA-2204-1743

Received date: 2022-04-27
Accepted date: 2022-06-14

Developing Financial Distress Prediction Models Based on Imbalanced Dataset: Random Undersampling and Clustering Based Undersampling Approaches

Seyed Behrooz Razavi Ghomi, Alireza Mehrazin*, Mohammadreza Shoorvarzi, Abolghasem Massihabadi

Ph.D. Student, Department of Accounting, Neyshabur Branch, Islamic Azad University, Neyshabur, Iran
Assistant Professor, Department of Accounting, Neyshabur Branch, Islamic Azad University, Neyshabur, Iran
Associate Professor, Department of Accounting, Neyshabur Branch, Islamic Azad University, Neyshaour, Iran
Assistant Professor, Department of Accounting, Sabzevar Branch ,Islamic Azad University, Sabzevar, Iran

ARTICLE INFO

Article history:
Received 00 December 00
Accepted 00 February 00

Keywords:
Imbalanced datasets,
Undersampling
financial distress prediction models,
financial ratios,
machine learning

ABSTRACT

So far, distress prediction models have been based on balanced, such sampling is not consistent with the reality of the statistical community of companies. If the data are balanced, the bias in sample selection may lead to an underestimation of typeI error and an overestimation of the typeII error of models. Although imbalanced data-based models are compatible with reality, they have a higher typeI error compared to balanced data-based models. The cost of typeI error is more important to Beneficiaries than the cost of typeII error. In this study, for reducing typeI error of imbalanced data-based models, random and clustering-based undersampling were used. Tested data included 760 companies since 2007-2017 with 4 different degrees and the results of the H1 to H3 test represented them. In all cases of the typeI error, typeII error of balanced data-based models were lower and more, respectively, compared to imbalanced data-based models; also, in most cases, the geometric mean of balanced data-based models was higher compared to imbalanced data-based models, respectively. The results of testing H4 to H6 show that in most cases, typeI error, typeII error and the geometric mean criterion of models based on modified imbalanced data were less, more, and more, respectiively compared to the models based on imbalanced data, in other words, applying Undersampling methods on imbalanced training data led to a decrease in typeI error and an increase in typeII error and geometric mean criteria. As a result using models based on modified imbalanced data is suggested to Beneficiaries

1.Introduction

Financial distress is a serious issue for the economic life of countries. The individual and social costs of financial distress have raised the issue of financial distress prediction for many managers, banks, investors, policymakers, and auditors. Financial distress prediction is of great importance to the three groups, including managers, creditors, and auditors. As the representatives of shareholders, managers pursue activities that seek continuity and profitability of the company. To assess the ability of a company to repay its obligations, lenders are willing to assess the continuity of business units. Auditors, as another of these groups, must comment on the financial statements, the continuity of the clients, and the fairness of the information in the financial statements. Therefore, they are interested in predicting financial distress or the continuation of companies[23]. So far, financial distress prediction models have been proposed with different variables and prediction methods [4]. Most of the proposed financial distress prediction models are designed using the traditional paradigm or

pairwise sampling ([11];[32]), meaning that the data include an equal number of healthy and financially distressed firms. But, such sampling does not correspond to the reality of the statistical community of companies. In cases where the number of financially distressed companies is low and the number of healthy companies is very high, researchers will encounter imbalanced data [2]. Zmijewski [42] showed that if the ratio or number of healthy firms to the proportion or number of financially distressed firms does not correspond to reality or the balanced data, sample bias may lead to underestimation of type I error and overestimation of type II error in financial distress prediction models. Financial distress prediction models based on imbalanced data will face the problem of higher type I error compared to balanced data. The cost of type I error is higher for investors, creditors, and other stakeholders than the cost of type II error in financial distress prediction models. Since type I error is more costly for investors and creditors than type II error, it is necessary to reduce type I error of financial distress prediction models based on imbalanced data. Given the mentioned points, not enough attention has been paid to the comparison of performance evaluation criteria (type I error, type II error, geometric mean): A: Financial distress prediction models based on imbalanced data (with different degrees of data imbalance) compared to balanced data-based models and B: Financial distress prediction models based on modified imbalanced data (with different degrees of data imbalance) compared to the models based on imbalanced data; so, research innovation compares performance evaluation criteria (type I error, type II error, geometric mean): A: Financial distress prediction models based on balanced data in comparison with models based on imbalanced data with 4 degrees, including 60% to 40%, 70% to 30%, 80% to 20% and 90% to 10%, B: Financial distress prediction models based on imbalanced data modified by undersampling, random undersampling and clustering-based methods in comparison with imbalanced data-based models with 4 degrees of 60% to 40%, 70% to 30%, 80% to 20% and 90% to 10% for solving the problem of imbalanced data in financial distress prediction (reducing type I error of financial distress prediction models). In this study, random undersampling and clustering-based undersampling were suggested to reduce type I error of financial distress prediction model based on imbalanced data with 4 different degrees [41]. The mentioned sampling methods lead to reducing the number of healthy companies in the training data sample and equalizing the number of healthy companies with the number of financially distressed companies in the training data. In such conditions, financial distress prediction models will be equally inclined to healthy companies and financially distressed companies, and type I error of financial distress prediction models is expected to be reduced, but the test data of the financial distress prediction models will be imbalanced according to reality.

The development of financial distress prediction models based on imbalanced data is necessary because the type I error and type II error of financial distress prediction models based on balanced data are less and more than the models based on imbalanced data-models in reality, respectively. So, models of financial distress prediction based on balanced data are not very suitable for decision-making of investors, creditors and other stakeholders in real conditions. type I error of financial distress prediction models based on imbalanced data is higher than the model based on balanced data[42] and the cost of type I error is higher than the cost of type II error for investors[15]. So, it is essential to reduce type I error of financial distress prediction models based on imbalanced data by using sampling and balanced training data.

It is recommended to all stakeholders to use financial distress prediction models based on modified imbalanced data in comparison with the financial distress prediction models based on balanced and imbalanced data; because, type I error is the measure of the geometric mean of the financial distress prediction models based on modified imbalanced data compared to models based on real balanced data and the models based on imbalanced data. Therefore, financial distress prediction models based on modified imbalanced data are expected to impose lower costs on stakeholders compared to imbalanced data. In addition, it is expected that by modifying the training process, the type I error and the geometric mean of the modified imbalanced data-based models will be less and more, respectively compared to theimbalanceddata-based models. Therefore, modified imbalanced data-based models are expected to impose lower costs on stakeholders compared to imbalanced data.

1.2.Theoretical Foundation

1.2.1.Definition and criteria for measuring financial distress

The term "early warning" is derived from the military field and is now used in other fields, such as macroeconomics, business management, environmental monitoring, finance, and others. Early warning about financial distress and bankruptcy is the subject of important research for corporate financing, the core of which is financial distress prediction. In general, financial distress prediction models through mathematical models, statistical models, and artificial intelligence models predict whether the company will suffer financial distress in the future based on financial data or not? Predicting distress plays an important role in management decisions for companies, investment decisions for investors, credit decisions for creditors, and credit ratings of banks,. Financial distress refers to a situation in which the company does not have enough cash flow to meet its financial obligations, and in such conditions, there will be serious consequences for the stakeholders. In such circumstances, managers make their decisions based on leaving the stage of financial distress faster and preventing the aggravation of financial distress and the occurrence of bankruptcy conditions [18]. Newton [31] divided the stages of a firm's unfavorable financial situation into a period of latency, deficit, inability to pay financial or commercial debts, inability to pay full debt and ultimately bankruptcy. Although most bankruptcies follow these steps, some companies may go bankrupt without going through all the steps.

In times of financial distress, companies face two main problems: Lack of liquidity in the balance sheet and the existence of many obligations. In other words, in times of financial distress, cash flows do not provide the necessary coverage to meet obligations and the company suffers a temporary inability to pay its debts. In this case, companies sell assets and receive loans, which results in reduced production capacity and performance and increased leverage. For this reason, predicting financial distress of companies is essential and provides the possibility of offering possible solutions before any crisis occurs [28].

1.2.2.Imbalanced Data

Imbalanced data exists if at least one of the values of a dependent variable significantly has a smaller sample than other values of the dependent variable [38]. When data are highly imbalanced ,the performance of prediction models is affected [21]; [8]; [3]). In the area of financial distress prediction, the imbalanced data scenario is due to limited samples in the minority class (financially distressed companies). Since the number of financially distressed companies is less than the number of healthy companies, financially distressed companies represent the minority class and healthy companies represent the majority class in imbalanced data. Imbalanced datasets generally increase the geometric Mean of model predictions. However, maximizing the geometric Meanmay not be the best approach for imbalanced data; maximizing the overall Geometric Meanof the model is due to the geometric Meanof the majority class (healthy companies); because, they have more weight in the selected sample. As a result of prediction models, they have a low error rate for the majority class (healthy companies) and a high error rate for the minority class (financially distressed companies), mainly type I error is more important than type II error. Because, the cost of the type I error is much higher than the cost of type II error. In the field of machine learning, the prediction model is more desirable to minimize the amount of type I error and the amount of type II error and maximize the geometric mean of the prediction model [21]. The main assumption of prediction models is balance or equality of the majority class(healthy companies) with the minority class (financially distressed companies). The weakness of prediction models based on imbalanced data is due to the learning phase of prediction models. That is, during the learning phase, prediction models tend to the geometric Mean of the majority class (healthy companies) and the Geometric Meanof the minority class (financially distressed companies) are ignored because the design of prediction models is such as to maximize the overall geometric Mean of prediction models [22]. Accordingly, it can be concluded that the prediction models in the training phase are inclined to the majority class (healthy companies); so, the models in the test phase accurately predict the majority class (healthy companies); as a result, type II error rate of imbalanced data-based prediction models is lower than that of balanced data-based models; but, the models often incorrectly predict the minority class (financially distressed companies); so, the rate of type I error

prediction models based on imbalanced data is more than the models based on balanced data. Thus, type I error of prediction models based onimbalanceddata originates from the learning stage of predictive models, where due to data imbalance, the performance of prediction models is disrupted. Prediction models are designed to optimize geometric Mean and do not take into account the relative distribution of each class [26].

1.2.3. Financial Prediction Models Based Imbalanced Data

Since the time of Beaver [5] and Altman [1], who first examined the prediction of financial distress, the traditional paradigm for financial distress prediction models is based on selected balanced data samples with available financial information in which the ratio of financially distressed firms equals the proportion of healthy firms; balanced samples can be created using sampling techniques known as pairwise samples [16]. This sample selection strategy has clear advantages as it prevents class errors in the learning phase. Classifiers maximize the geometric Mean of the prediction model regardless of the class distribution. However, this strategy has a serious problem that does not reflect the statistical population distribution of healthy and financially distressed companies in reality. Zmijewski [42] showed that if the ratio of healthy companies to financially distressed companies does not reflect the statistical population of companies in reality, an error or bias occurs in the selecting the sample of companies, which leads to underestimation of the type I error and overestimation of type II error. In addition, Ooghe and Joos [33]argued that the sample of healthy and financially distressed firms should represent the statistical population of firms in order to use financial distress prediction models.

Mostly researchers have used balanced data samples with equal numbers of healthy companies and financially distressed companies to predict financial distress; however, according to Zmijewski [42], if the ratio of financially distressed firms to healthy companies is different from reality, the prediction power of the models will be impaired and in particular, there will be a negative relationship between the number or ratio of financially distressed companies with type I error rate in the sense that by reducing the number or ratio of financially distressed companies, the type I error rate increases (type I error means that the model classifies a financially distressed company as a healthy company); also, there is a positive relationship between the number or ratio of financially distressed companies with type II error rate. This means that by reducing the number or proportion of financially distressed companies, type II error rate decreases (type II error means that the model of a healthy company is misclassified as a financially distressed company). Therefore, if the number or ratio of financially distressed companies is greater than the reality (balanced data), this will lead to underestimation of type I error and overestimation of type II error and vice versa. Another issue to consider is that the ratio of healthy firms to financially distressed firms varies in reality. So, evaluating the performance of financial distress prediction models based on a range of imbalanced data with varying degrees is necessary.

1.2.4.Modified Imbalanced Data-Based Financial Prediction Models

If the ratio of financially distressed to healthy companies differs from reality, the prediction power of the models will be impaired [42]; in other words, type I error of financial distress prediction models based on imbalanced data is more than balanced data; since type I error imposes more costs on investors, creditors, and other stakeholders than the type II error, it is necessary to reduce type I error of models based on imbalanced data to reduce the costs imposed on stakeholders. Literally, the way to reduce type I error of financial distress prediction models based on imbalanced data is called solving the problem of imbalanced data in financial distress prediction. In previous studies, in order to reduce type I error of financial distress prediction models based on imbalanced data (problem solving of imbalanced data), 3 approaches at the data level, at the algorithm level and cost-sensitive learning were used. Data level solution includes open sampling methods, the most important of which is undersampling, oversampling, and combined method; another solution at the algorithm level includes modified algorithm and one-tier learning [9]. Data-level sampling methods are performed independently of the prediction models. Sampling methods can be combined with other prediction models. Sampling methods, including undersampling, oversampling and combined method lead to converting primary

imbalanced data to balanced data. Balanced data by sampling techniques improve the performance of prediction models [13]. Undersampling in general is a method to exclude samples from the majority class, and in particular in the area of financial distress prediction is a method to exclude healthy companies from the selected sample in order to create balance and equality with the minority class or companies. The advantage of the oversampling method is the preservation of the natural distribution of the minority class (financially distressed companies) without omitting the initial data. But, the disadvantage is that it artificially increases the data of the minority class (financially distressed companies) and can lead to performance disruption and reduction of the prediction models. Undersampling is an efficient method in sampling as it uses relatively less data but the risk is to remove useful data from the machine learning process.

1.3. Empirical Background

McKee and Greenstein(2000) [29] examined the ability of 3 models to predict financial distress for 5 sets of highly imbalanced training data and test data. The ratio of financially distressed companies to healthy companies reflected the reality of the statistical population of companies. Results showed that the imbalanced distribution of samples in the learning process has led to poor performance of classification models, especially for financially distressed companies. Raei and Fallahpour(2004)[34] predicted the financial distress of manufacturing companies using artificial neural networks. Results showed that the artificial neural network model is significantly more accurate in financial distress prediction than the multiple discrimination analysis model. Saruei(2010) [37] examined the revenue of Springit, Zimsky, and Olsen models in the pharmaceutical and textile industries. Their results showed that in all three years, the used models in the textile industry outperformed the others. Veganzones and Severin(2018) [40]conducted a combined study on the performance reduction of financial distress prediction models based on imbalanced data with varying degrees and performance improvement methods (i.e. solving imbalanced data problem) of corporate financial distress prediction models. Given that the performance of financial distress prediction models based on imbalanced data is lower compared to balanced data, determining the ability of imbalanced data problem-solving methods and the improvement rate of performance of financial distress prediction models based on modified imbalanced data are important. Results of this study showed that an imbalanced distribution in which the minority class is 20% of the total number of selected companies significantly disrupts the performance of imbalanced data-based financial distress prediction models. In addition, support vector machine model is less sensitive than other models of financial distress prediction based on imbalanced data, and sampling methods can improve the performance of financial distress prediction models based on imbalanced data.Ghasemi and Sarlak [14] examined the impact of the financial crisis on the financial transparency and conservatism in the banking industry. To this end, they collected data from 18 banks since 2011-2015. The results of a linear regression analysis showed the correlation and the impact of the financial crisis on the financial transparency and conservatism. Zoricak et [43] predicted the bankruptcy of small and medium-sized enterprises with imbalanced data (with varying degrees). The data showed that the real ratio between healthy companies and financially distressed companies is highly imbalanced . The solution of the mentioned research to overcome the problem of imbalanced data (imbalanced learning) is to use one-tier prediction models. In the mentioned research, isolated forest technique and one-tier support vector machine were used. Results showed the best score of geometric mean to be 91%. Rezaei and Javaheri [36] compared the neural network's predictability with the combined method of the genetic algorithm and artificial neural network. For this purpose, they selected a sample of 58 healthy companies and 49 financially distressed companies. In order to compare these methods, determination coefficient, MSE and RMSE were used. The results showed 97.7% accuracy of the artificial neural network and 100% accuracy of the combined method of the artificial neural network and genetic algorithm in predicting healthy and financially distressed companies. Therefore, the combined method of artificial neural network and genetic algorithm was the best way to predict the financial crisis of the companies.Haghparast et al [17] predicted the bankruptcy of companies using convolutional neural network and visual financial ratios. The research period was from 2009 to 2018 and the sample companies were listed

on the Tehran Stock Exchange in two groups, including 66 healthy companies and 66 bankrupt companies. The analysis results showed that the convolution neural network model had the ability to predict with 50% accuracy**.** Razavi et al [35] compared the performance of financial distress prediction models based on imbalanced data with varying degrees from 2007 to 2017. For this purpose, logistic regression models, multiple discrimination analysis, support vector machine with sigmoid kernels, random forest, nearest neighbor algorithm, and artificial neural network were used. To determine the optimal parameters, the combined method of grid search and cross-sectional validation was used. The results showed that the best performance for balanced and imbalanced data with lower degrees was related to the random forest model while for imbalanced data with higher degrees, it was related to the support vector machine model with sigmoid, radial, and linear kernel functions. Also, the performance of the nearest neighbor algorithm was not significantly different from other models; besides, the performance of linear models, such as linear discrimination analysis and logistic regression were weaker than nonlinear models

1.4. Hypotheses

H1. Type I error (misidentification of a financially distressed company as a healthy company) in the models of financial distress prediction based on balanced data is significantly less than the models based on imbalanced data.

H2. Type II error (misidentification of a healthy company as a financially distressed company) of financial distress prediction models based on balanced data is significantly more than models based on imbalanced data.

H3 .The geometric mean of financial distress prediction models based on balanced data is significantly higher than models based on imbalanced data.

H4 .Type I error (misidentification of a financially distressed company as a healthy company) in models of financial distress prediction based on modified imbalanced data is significantly less than models based on imbalanced data.

H5 .Type II error (misidentification of a healthy company as a financially distressed company) in models of financial distress prediction based on modified imbalanced data is significantly more than the models based on imbalanced data.

H6. The geometric mean of financial distress prediction models based on modified imbalanced data is significantly higher than the models based on imbalanced data.

2. Methodology

This research was applied. The research design was quasi-experimental, using post-event approach. In order to test the hypotheses, 4 data sets, including 760 companies with different degrees of 60% to 40%, 70% to 30%, 80% to 20% and 90% to 10% were used. In order to createimbalanceddata with the mentioned degrees, the method of Brown and Mues [7] was used. For each selected company, 64 financial ratios were collected as a predictor variable of financial distress using the information in the audited financial statements of the company. To test the research hypotheses, two financial distress prediction models, including support vector machine with radial function and random forest were used. Since parameter adjustment plays a vital role in the performance of financial distress prediction models, network search optimization method was used to optimize the parameters. Also, cross-validation with number 5 was used to increase the reliability of financial distress prediction models. In this way, in each implementation of the financial distress prediction model, 80% of the data were selected as training data and the remaining 20% were selected as test data. This operation was performed 5 times for each set of data so that all data were tested. In order to compare the differences between the performance evaluation criteria significantly, Mann-Whitney test [27] was used at 90% and 95%confidence levels.The statistical population of the study included all companies listed in Tehran Stock Exchange. Companies with the following 3 conditions were selected as a sample:

1.Selected companies in all research periods should be in the list of companies listed in Tehran Stock Exchange.

2.The end of the fiscal year of the selected companies in all research periods should be unchanged and during the research period, the information required by the selected companies should be available.
3.The selected companies are not in the list of investment companies, banks and financial intermediation institutions; this group was excluded from the sample due to different financial structure (high use of financial leverage). Selected sample data were collected from the financial statements of companies listed in the Tehran Stock Exchange since 2007-2017. In this research, for testing the hypotheses, 4 data sets including 760 companies with different degrees of 60% to 40%, 70% to 30%, 80% to 20%, 90% to 10% were used. In order to create imbalanced data with the mentioned degrees, the method of Brown and Mues [7] was used; first, 380 healthy companies and 380 financially distressed companies were randomly selected and then, 380 primary financially distressed companies were randomly removed and in the same number, new healthy companies were randomly added so that the total number of companies for five sets of balanced and imbalanced data gets always equal to 760 companies. Table 1 shows the sample distribution of balanced and imbalanced data by healthy and financially distressed companies.

Table 1. Distribution of balanced and imbalanced data samples based on healthy and financially distressed companies

Test set size	Size of training set	Number of financially distressed companies	Number of healthy companies	Total No of companies	Ratio of healthy to financially distressed companies
152	608	380	380	760	%50 VS %50
152	608	304	456	760	%60 VS %40
152	608	228	532	760	%70 VS %30
152	608	152	608	760	%80 VS %20
152	608	76	684	760	%90 VS %10

2.1.Dependent Variable

The dependent variable of the research was determined according to Article 141 of the Commercial Code; so that if the accumulated loss of the company was more than 50% of the company's capital, the company would be financially distressed and dependent variable took the value of 1; otherwise, the company would be healthy and the dependent variable took the value of 0.

2.2. Independent or Predictive Variables

After reviewing the literature on the predictor variables of companies' financial distress, 64 prediction variables were selected. Table 2 shows the list and how to calculate the 64 predictive variables. It shows the financial ratios used to predict financial distress.

Table 2. The financial ratios used to predict financial distress

Row	Variable	Row	Variable	Row	Variable	Row	Variable
1	NI/SE	20	S/FA	39	CL/TL	57	RE/Inv
2	NI/TA	21	S/SE	40	D/NI	58	RE/SC
3	OCF	22	S/TA	41	EPS	59	RE/TA
4	OCF/SE	23	SE/TA	42	EBIT/IE	60	S/Ca
5	OCF/CL	24	SE/TL	43	EBIT/S	61	CA/S
6	OCF/IE	25	Size(log TA)	44	EBIT/TA	62	CA/TA
7	OCF/S	26	TIBL/TL	45	FA/(SE+LTD)	63	CL/SE
8	OCF/TA	27	TL/TA	46	FA/TA	64	CL/TA
9	OCF/TL	28	WC/S	47	GP/S		
10	OCF/NI	29	WC/TA	48	IE/GP		
11	OCF/OI	30	(Ca+STI)/CL	49	IE/S		
12	NI/GP	31	R+Inv)/TA	50	Inv/WC		
13	OI/S	32	P/S	51	Inv/S		
14	OI/TA	33	R/S	52	LTD/SE		

15	PIC/SE	34	Ca/CL	53	LTD/TA		
16	QA/CL	35	Ca/TA	54	MVE/TA		
17	QA/Inv	36	NI/S	55	MVE/TL		
18	QA/TA	37	CA/CL	56	MVE/SE		

Where, CA is current assets, NI is net profit, Ca is cash income, OI is operating income, CL is current liquidity, QA is current assets, PIC is paid capital, R is receivables, EBIT is earnings before interest and taxes, RE is return on equity, FA is fixed assets, S is revenues, GP is gross profit, SC is stock capital, IE is income expenses, SE is stock equity, INV is inventory, STI is short term investment, TA is total assets, LTD is long term debts, TL is total liquidity, MVE is market value equity, WC is working capital, OCF is operating cash flow, D is dividend, TIBL is Total interest on debt.

2.3. Experimental Framework for Imbalanced Data Undersampling

Undersampling will reduce the number of samples in the majority of imbalanced data and balanced data. In this study, in financial distress prediction based on imbalanced data, healthy companies represent the majority of imbalanced data and financially distressed companies represent the minority of imbalanced data. Then, undersampling methods were used to reduce the number of healthy companies in imbalanced training data. Reducing the number of healthy companies inimbalanced training data led to balancing the number of healthy companies and the number of financially distressed companies in training data. Figure1 shows how to implement data undersampling methods for training data as input to financial distress prediction models:

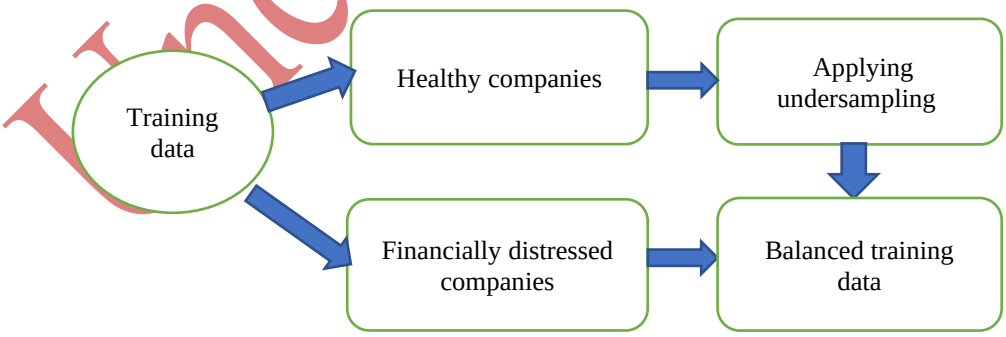

Fig 1. Experimental framework for undersampling training data

2.4.Random Undersampling Method

Random undersampling is a method to randomly reduce the samples in the majority class (healthy companies) ofimbalanceddata to balance the number of samples in the majority class (healthy companies) with the number

of samples in the minority class (financially distressed companies). In this study, for applying the undersampling method, first, the input training data set of financial distress prediction models was divided into two majority (healthy companies) and minority classes (financially distressed companies); then, using random undersampling method , a number of majority class samples (healthy companies) was randomly selected and deleted so that the number of majority class samples (healthy companies) equalized with the number of minority class samples (financially distressed companies). In the next step, the samples of the minority class (financially distressed companies) were combined with the samples of the majority class (healthy companies) and balanced training data were created as inputs to financial distress prediction models.

2.5.Clustering-Based Undersampling Method

Clustering-based undersampling method is used to reduce the number of samples in the majority class (healthy companies) using machine learning and is among the non-supervisory learning methods. In this study, for applying undersampling methods, first, as inputs of financial distress prediction models, training data was divided into majority class (healthy companies) and minority class (financially distressed companies); then, using clustering-based undersampling method, the number of majority class samples (healthy companies) was reduced so that the number of majority class samples (healthy companies) got equal to the number of minority class samples (financially distressed companies); in the next stage, minority class samples (financially distressed companies) were combined with majority class samples (healthy companies) and balanced training data were generated as input to financial distress prediction models. Although there are various clustering-based algorithms, in this research, K-Medoids clustering algorithm was used to reduce the number of majority class samples (healthy companies); since, K-Medoids clustering algorithm has been widely used in other studies [41]. In this method, the number of cluster centers is set equal to the number of minority class samples (financially distressed companies) (K = N). Therefore, the clustering algorithm includes K cluster centers of all majority class samples (healthy companies), cluster centers are examples of the same cluster (healthy companies) and will replace the majority class samples (healthy companies); finally, the number of majority class samples (healthy companies) will be equal to the number of minority class samples (financially distressed companies)[41]. In this study, the Euclidean distance criterion was used to implement the K-Medoids clustering algorithm.

2.6.Financial Distress Prediction Models

2.6.1.Random Forest

Random forest is a collective decision tree in which each classifier is created using a random vector independent of the input vector [6]. In the random forest prediction model, each tree is created from a bootstrap sample of the decomposed data. At the end, the classification is determined by a majority of votes for each item by aggregating the classification tree. When creating a tree, the random forest searches for a random subset of input features in the decomposition of each node, and the tree is allowed to grow completely without pruning, because only part of the input features are used and pruning is not performed. Random forest has quick and simple calculations and good performance**.**

2.6.2.Least Squares Support Vector Machine

Support vector machine is one of the supervisory learning techniques used in classification and regression. Its basic principles are the construction of a separator with a maximum margin in the feature space; instead of an exact conversion, the principle of substituting the kernel function was used to convert it to a nonlinear model. Vapnik [39] proposed the least squares support vector machine (LS-SVM) in order to further adapt to the original support machine equation, which led to the solution of linear system of the Karush, Cohen, Tucker (instead of the problem of more complex second order planning):

MIN w .b. e $=\frac{1}{2} w^t w + c \sum_{i=1}^{N} ei$
Subject to yi (w Ω(xi)+b)+ei-1≥0 ei≥0
$$f(y) = sign(\sum_{i=1}^{N} yipik(x, xi) + b \quad (1)$$

2.7.Criteria for Evaluating the Performance of Financial Distress Prediction Models

In this study, for comparing the performance of financial distress prediction models based on balanced data, imbalanced data and modified imbalanced data, the performance evaluation criteria of type I error, type II error, and geometric mean were used.

2.8.Criteria for Type I Error and Type II Error

For calculating type I error and type II error rate, the confusion matrix was used (Hesam et al., 2020).

Table 3. The confusion matrix

Predicted negative	Predicted positive	Confusion matrix
False negative (FN)	True positive (TP)	True positive
True negative (TN)	False positive (FP)	True negative

Where, TPR is percentage of positively classified samples, also known as sensitivity; in the area of financial distress prediction, it means the percentage of financially distressed companies that are correctly classified as financially distressed companies [40]. So, the method to calculate the sensitivity and type I error was described in Equation 2 and 3:

$$TPR = Sensitivity = \frac{TP}{TP + FN} \quad (2)$$

$$Erorr\ Type\ \text{I} = 1 - \frac{TP}{TP + FN} \quad (3)$$

Where, TNR is percentage of negatively classified samples, also referred to as specificity. In the field of predicting financial distress, it refers to the percentage of healthy companies that are correctly classified as healthy companies [40]; so, the method to calculate the specificity and type II error is described in Equation 4 and 5.

$$TNR = Specificity = \frac{TN}{TN + FP} \quad (4)$$

$$Erorr\ Type\ \text{II} = 1 - \frac{TN}{TN + FP} \quad (5)$$

2.9.Geometric Mean

Geometric mean is a reliable criterion in assessing the predictability of financial distress prediction models based onimbalanceddata; because, in calculating it, the majority class (healthy companies) and the minority class (financially distressed companies) are considered, respectively. Based on Equation.6, the geometric mean includes the Geometric Mean of predicting financially distresssed companies (sensitivity) and the correct prediction Geometric Meanof healthy companies (specificity). Financial distress prediction model is capable to predict or classify high percentage of healthy and financially distressed companies at the same time and thus has a high geometric mean. A higher geometric mean indicates better performance of the prediction model for two classes and no bias of the prediction model towards one class. In Equation 6, TPR is the percentage of financially distressed companies that the model is correctly placed in the category of financially distressed companies and TNR is the percentage of healthy companies. So, the model is properly placed in the healthy class [23]. The method to calculate general Geometric Mean by geometric mean is described in Equation 6:

$$G - mean = \sqrt{TPR * TNR} \quad (6)$$

2.10.Parameter Optimization with a Combined Method of Grid Search and Cross-Validation

Grid search is a representation of search based on a defined set of optimal parameter space (hyperparameters) [24]. The reasons for using grid search optimization method are as follows: There may not be enough certainty about meta-heuristic optimization methods since approximate and meta-heuristic optimization methods prevent the complete search of parameters. Another reason is lower computational time in finding the values of optimal parameters compared to other more advanced methods [25]. In the grid search optimization method, the optimal parameters are determined using the minimum value (lower limit), the maximum value (upper limit), and the number of steps.

2.11.Tuning Parameters

In this paper, the radial function of the least squares support vector machine was used, which included two parameters of c and gamma. The gamma and c parameters have a vital role in the performance of the least squares support vector machine model [19]. Therefore, improper selection of them can lead to overestimation and underestimation problems. However, there are few practical guidelines in determining the optimal parameters. Hsu et al [19] suggested a practical guide for tuning parameters of least squares support vector machine using combined method of grid search and cross-validation. The purpose of adjusting c and gamma parameters is to maximize the Geometric Mean of the prediction model for unseen data. Exponential growth of c and gamma parameters is a practical method in tuning parameters. In this study, the values of 2^{-5}, 2^{-3}, 2^{-1}, 2, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11} , 2^{13}, 2^{15} for the parameter c and the values of 2^{-15}, 2^{-13}, 2^{-11}, 2^{-9}, 2^{-5}, 2^{-7},2^{-3} ,2 , 2^{3}, and 2^{5} have been used for the gamma parameter; so, a total of 121 combinations of parameters have been selected and for determining the optimal parameter of the random forest prediction model with the criterion of Gain_ Ratio, 28 values of the number of trees parameter were tuned in the range of 10 to 1000.

2.12.Experimental Framework for Testing Hypotheses

In this research, according to Figure 2, in order to calculate the performance evaluation criteria of financial distress prediction models based on balanced data, imbalanced data, and testing H1 to H3, the following steps were followed. First, the data were divided into two training and test categories. 80% of the data were selected as training data and 20% as test data. Financial distress prediction models were trained separately with balanced training data and imbalanced training data with 4 different degrees. In step 2, the optimal parameters of financial distress prediction models were determined. In stage 3, the financial distress prediction models trained in stage 2 were evaluated with balanced test data and imbalanced data with 4 different degrees. In the last step, the performance evaluation criteria of financial distress prediction models, including type I error, type II error, and the geometric mean of the general Geometric Mean criterion were calculated. In stage 3, the financial distress prediction models trained in stage 2 were evaluated with balanced test data and imbalanced data with 4 different degrees. In the last step, the performance evaluation criteria of the financial distress prediction models including type I error, type II error and the geometric mean of the geometric mean criterion were calculated.

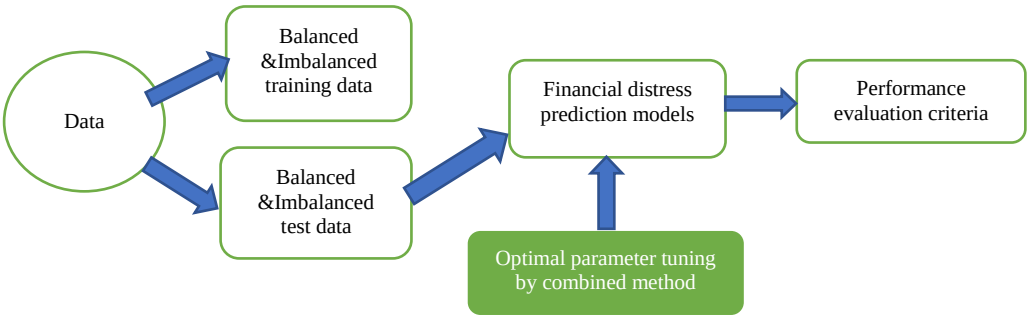

Fig 2. Experimental Framework for Testing H1- H3

According to Figure 3, in order to test the H4-H6 and calculate the performance evaluation criteria of financial distress prediction models based on modified imbalanced data, the following steps were performed:

1. Initially, the data were divided into training data and test data.
2. Training data on financial distress prediction models were divided into majority class (healthy companies) and minority class (financially distressed companies).
3. At this stage, using undersampling methods, the number of healthy companies in training data was reduced and equalized to the number of financially distressed companies in training data.
4. After applying undersampling in step 3, healthy companies were combined with financially distressed companies and balanced training data were obtained.
5. At this stage, financial distress prediction models were trained with balanced training data; in order to determine the optimal parameters, the combined grid search optimization method with cross-validation with the number 5 was used.
6. Finally, the performance evaluation criterion of financial distress prediction models based on modified imbalanced test data were calculated.

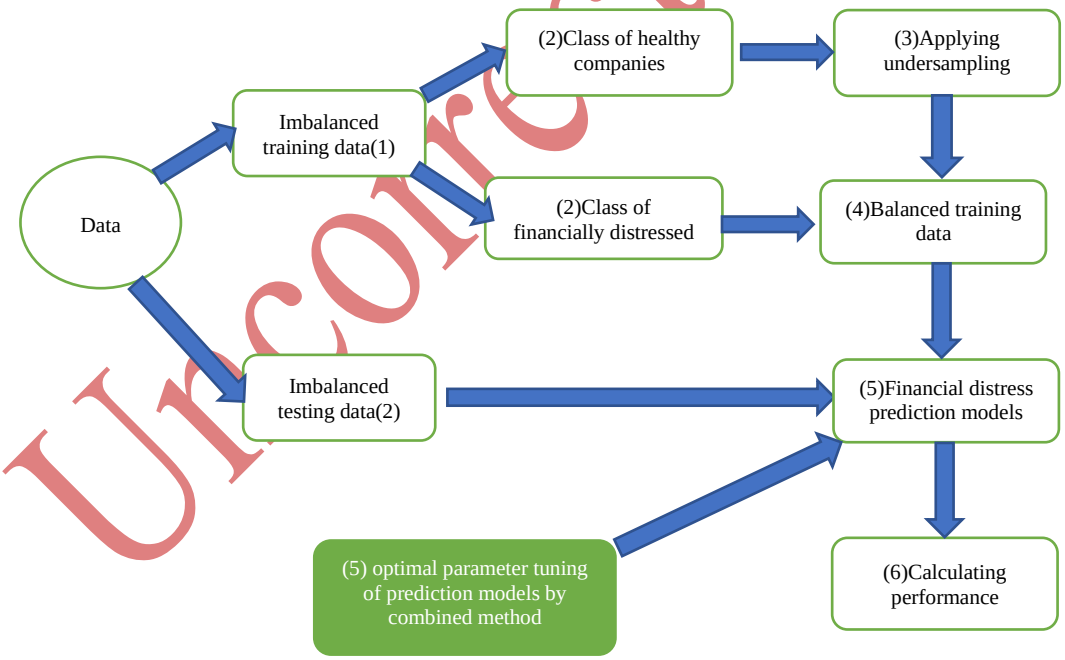

Fig 3. Experimental Framework for Testing H4 to H6

3. Results

H1 test result: Type I error (misidentification of a financially distressed company as a healthy company) in financial distress prediction models based on balanced data are less significant than the models based on

imbalanced data. According to Table 4 and Figure 4 and 5, type I error of financial distress prediction models based on balanced data is lower compared to imbalanced data-based models with 4 different degrees, including 60% to 40%, 70% to 30% ,80% to %20 and 90% to 10%. As expected, with the imbalanced data (increase in the number of healthy companies and decrease in the number of financially distressed companies), type I error of financial distress prediction models had an upward trend. Type I error of support vector machine based on balanced data was significantly lower in comparison with imbalanced data with degrees of 70% to 30% and 90% to 10% at 95% confidence level and imbalanced data with degrees of 60% to 40%, 70% to 30% and 90% to 10% at the 90% confidence level. Type I error of random forest model based on balanced data is significantly lower than imbalanced data with degrees of 70% to 30%, 80% to 20% and 90% to 10% at 95% and 90% confidence levels.

Table 4. Type I error comparison of financial distress prediction models based on balanced data with imbalanced data

Financial distress prediction models	Ratio of imbalanceddata	Type I error		Mann-whitney P-value	Hypotheses test result	
		Imbalanced data	Balanced data		95% confidence level	90% confidence level
SVM	%60 VS %40	%15.79	%10.66	0.095	Rejected	Confirmed
SVM	%70 VS %30	%22.84	%10.66	0.008	Confirmed	Confirmed
SVM	%80 VS %20	%30.05	%10.66	0.144	Rejected	Rejected
SVM	%90 VS %10	%38.32	%10.66	0.008	Confirmed	Confirmed
Random forest	%60 VS %40	%17.71	%15.16	0.403	Rejected	Rejected
Random forest	%70 VS %30	%28.48	%15.16	0.008	Confirmed	Confirmed
Random forest	%80 VS %20	%45.26	%15.16	0.008	Confirmed	Confirmed
Random forest	%90 VS %10	%53.79	%15.16	0.008	Confirmed	Confirmed

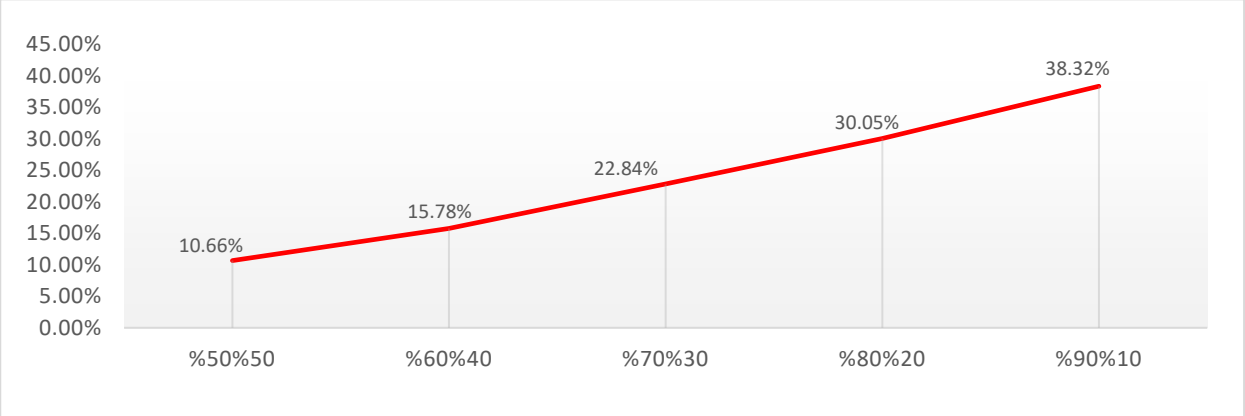

Fig 4. Type I error of SVM based on balanced data and imbalanced data

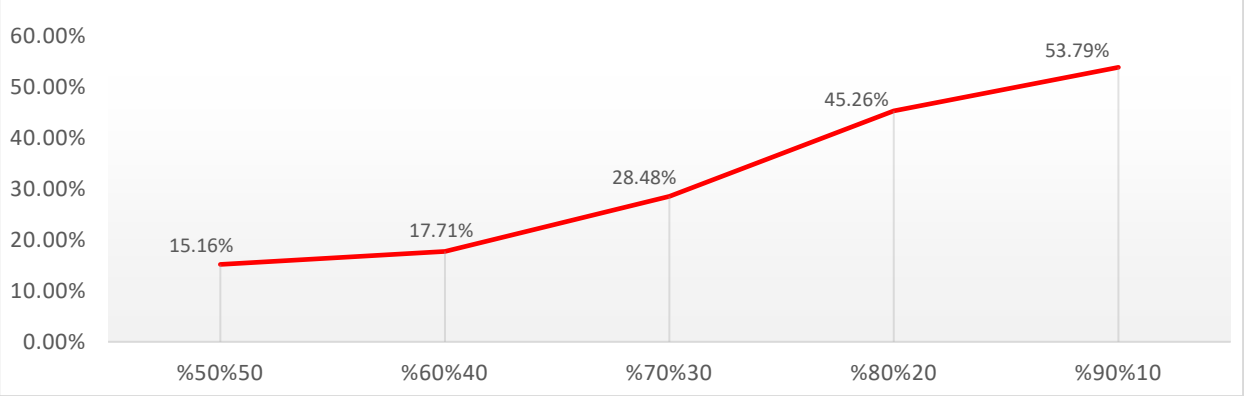

Fig 5. Type I error of Random Forest based on balanced data and imbalanced data

H2 test result: According to Table 5 and Figure 6 and 7, type II error of financial distress prediction models based on balanced data was higher compared to the imbalanced data-based models with 4 different degrees including 60% to 40%, 70% to 30%, %80 to 20% and 90% to 10%. As expected, with the imbalanced data (increased number of healthy companies and decreased number of financially distressed companies), type II error of financial distress prediction models had a downward trend. Besides, by increasing intensity of imbalanced data, type II error of financial distress prediction models based on balanced data had an upward trend compared to imbalanced data. According to H_2, the type II error of support vector machine model based on balanced data is significantly higher than the imbalanced data with 4 different degrees at 95% and 90% confidence levels, and the type II error of the random forest model based on balanced data is significantly higher than imbalanced data in A and B. Imbalanced data A:Imbalanced data with degrees of 70% to 30%, 80% to 20% and 90% to 10% at 95% confidence level and imbalanced data B:imbalanced data with degrees of 60% to 40%, 70% to 30%, 80% to 20% and 90% to 10% at 90% confidence level.

Table 5: Comparison of type II error of financial distress prediction models based on balanced data with imbalanced data

Financial distress prediction models	Ratio of imbalanced data	Type II error		Mann-whitney P-value	Hypotheses test result	
		Imbalanced data	Balanced data		95% confidence level	90% confidence level
SVM	%60 VS %40	%6.56	%14.76	0.008	Confirmed	Confirmed
SVM	%70 VS %30	%6.83	%14.76	0.008	Confirmed	Confirmed
SVM	%80 VS %20	%4.10	%14.76	0.012	Confirmed	Confirmed
SVM	%90 VS %10	%1.6	%14.76	0.008	Confirmed	Confirmed
Random forest	%60 VS %40	%4.39	%7.88	0.06	Rejected	Confirmed
Random forest	%70 VS %30	%1.88	%7.88	0.022	Confirmed	Confirmed
Random forest	%80 VS %20	%0.33	%7.88	0.011	Confirmed	Confirmed
Random forest	%90 VS %10	%0.44	%7.88	0.010	Confirmed	Confirmed

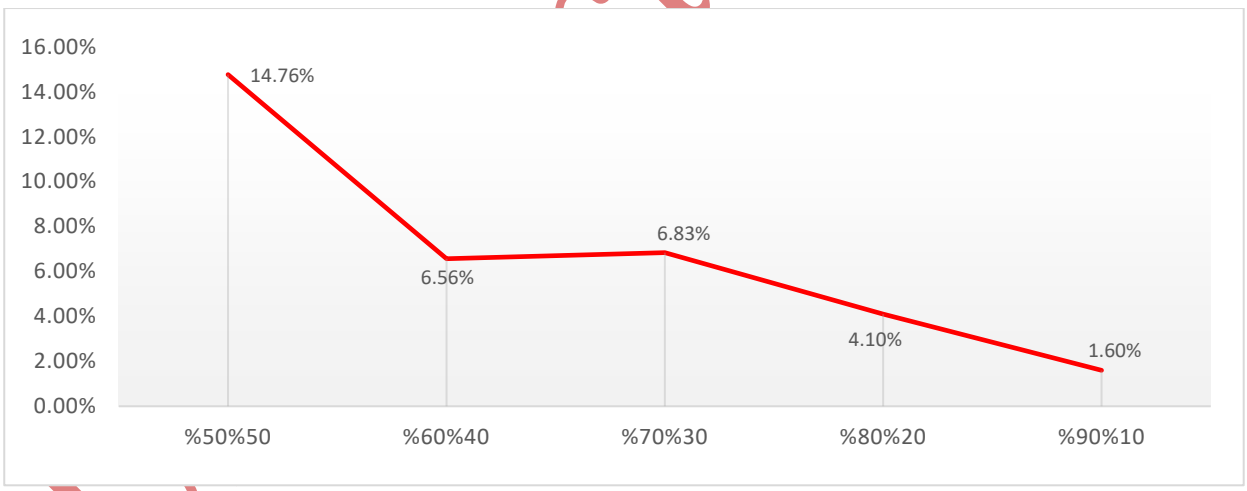

Fig 6. Type II error of SVM based on balanced and imbalanced data

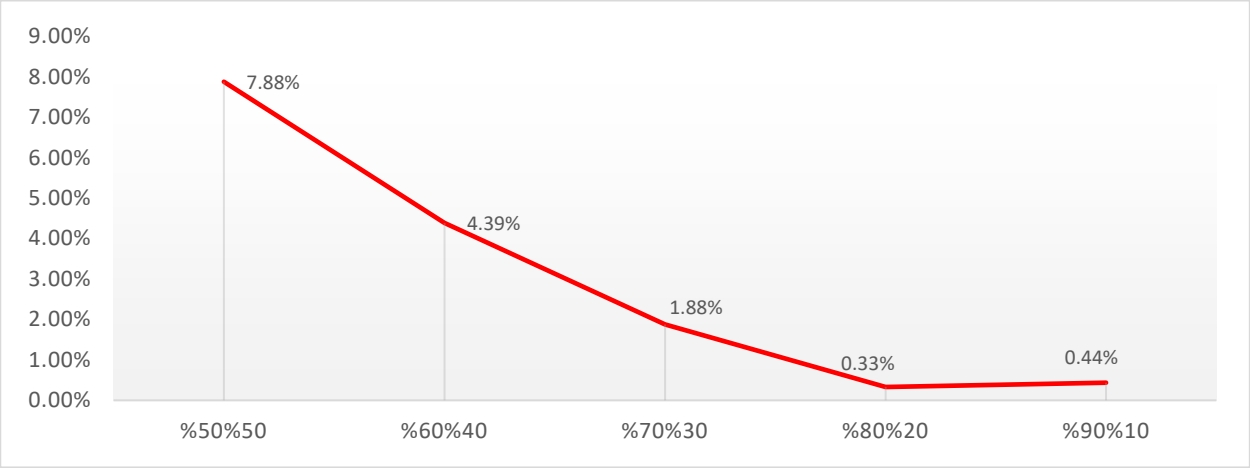

Fig 7. Type II error of Random Forest based on balanced and imbalanced data

H3 test result: Although H3 test results indicated that the geometric mean of the support vector machine model based on balanced data is higher compared to imbalanced data with degrees of 70% to 30%, 80% to 20% and 90% to 10%. The results of Mann-Whitney [26] test indicated that the geometric mean of the support vector machine model based on balanced data is significantly higher than the 90% to 10% imbalanced data at the 90% confidence level. Also, H3 implying that the random forest model based on balanced data is more accurate compared to imbalanced data with degrees of 70% to 30%, 80% to 20% and 90% to 10% at 95% and 90% confidence levels was confirmed.

Table 6. Comparison of Geometric Mean of financial distress prediction models based on balanced with imbalanced data

Financial distress prediction models	Ratio of imbalanced data	Geometric Mean		Mann-whitney P-value	Hypotheses test result	
		Imbalanced data	Balanced data		95% confidence level	90% confidence level
SVM	%60 VS %40	%88.67	%87.26	0.403	Rejected	Rejected
SVM	%70 VS %30	%84.66	%87.26	0.296	Rejected	Rejected
SVM	%80 VS %20	%81.51	%87.26	0.144	Rejected	Rejected
SVM	%90 VS %10	%76.26	%87.26	0.095	Rejected	Confirmed
Random forest	%60 VS %40	%88.68	%88.34	0.835	Rejected	Rejected
Random forest	%70 VS %30	%83.73	%88.34	0.037	Confirmed	Confirmed
Random forest	%80 VS %20	%73.57	%88.34	0.008	Confirmed	Confirmed
Random forest	%90 VS %10	%67.12	%88.34	0.008	Confirmed	Confirmed

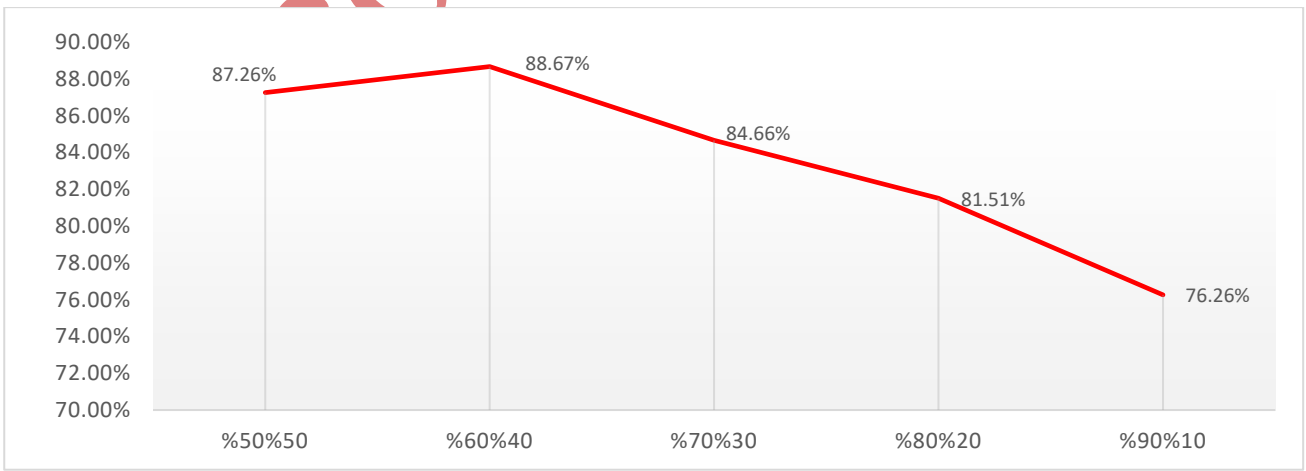

Fig 8. Geometric mean of support vector machine based on balanced and imbalanced data

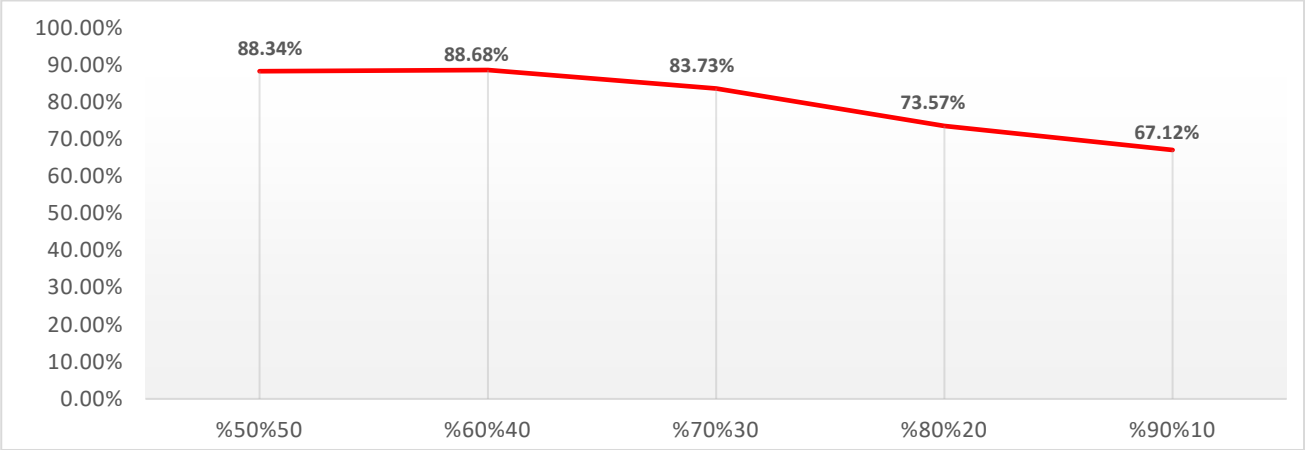

Fig 9. Geometric mean of Random Forest based on balanced and imbalanced data

H4 test result: According to Table 7, type I error (misidentification of a financially distressed company as a healthy company) in financial distress prediction models based on modified imbalanced data was lower compared to models based on imbalanced data. As seen in Figures 10- 13, as the data become more imbalanced, the blue line has an upward trend in all graphs. In other words, as the data become more imbalanced , type I error of financial distress prediction models based on imbalanced data has an upward trend.

According to Table 7 and Figures 10 -11, type I error of the support vector machine model based on modified imbalanced data by clustering-based undersampling method and random undersampling was lower compared to imbalanced data with 4 different degrees. Type I error of the support vector machine model based on modified imbalanced data (by clustering-based undersampling method) was significantly lower compared to imbalanced data with degrees of 70% to 30% and 90% to 10% at 95% confidence level and imbalanced data with 4 different degrees at the 90% confidence level. Type I error of the support vector machine based on modified imbalanced data by random undersampling method were significantly lower compared to imbalanced data with degrees 70% to 30% and 90% to 10% at 95% and 90% confidence levels. Therefore, in all cases, applying clustering-based undersampling and random undersampling on training imbalanced data led to reducing type I error of support vector model, which in most cases, reduced type I error of support vector machine model based on modified imbalanced data compared to models based on imbalanced data at 90% confidence level.

According to Table 7 and figures 12 to 13, type I error of random forest model based on modified imbalanced data by clustering-based undersampling and random undersampling was lower compared to imbalanced data with 4 different degrees. Type I error of random forest model based on modified imbalanced data by clustering-based undersampling method was significantly lower compared to imbalanced data with degrees of 70% to 30%, 80% to 20% and 90% to 10% at 95% confidence level andimbalanced data with 4 different degrees at 90% confidence level. Also, type I error of random forest model based on modified imbalanced data by random undersampling method was significantly lower compared to imbalanced data with 70% to 30%, 80% to 20% and 90% to 10% degrees at 95% and 90% confidence levels. Therefore, in all cases, applying clustering-based undersampling and random undersampling on training imbalanced data led to a reduction in type I error of random forest model. In all cases except for imbalanced data with 60% to 40, the reduction of type I error of the random forest model based on the modified imbalanced data was more significant in comparison with the models based on the imbalanced data at the 95% confidence level.

Table 7: Comparison of type I error of models based on balanced data in comparison with modified imbalanced data

Financial distress prediction model	Ratio of imbalanced data	Type I error of modified imbalanced data		Type I error of imbalanced data	P-value (clustering method)	P-value (random method)
		random	clustering			
SVM	%60 VS %40	%11.46	%9.88	%15.79	0.095*	0.296
SVM	%70 VS %30	%12.13	%13.87	%22.84	0.012**	0.008**
SVM	%80 VS %20	%13.37	%11.08	%30.05	0.095*	0.144
SVM	%90 VS %10	%11.24	%13.66	%38.32	0.037**	0.022**
Random forest	%60 VS %40	%12.09	%12.83	%17.71	0.095*	0.403
Random forest	%70 VS %30	%11.55	%11.61	%28.48	0.008**	0.008**
Random forest	%80 VS %20	%12.77	%12.83	%45.26	0.008**	0.008**
Random forest	%90 VS %10	%8.36	%11.12	%15.79	0.008**	0.012**

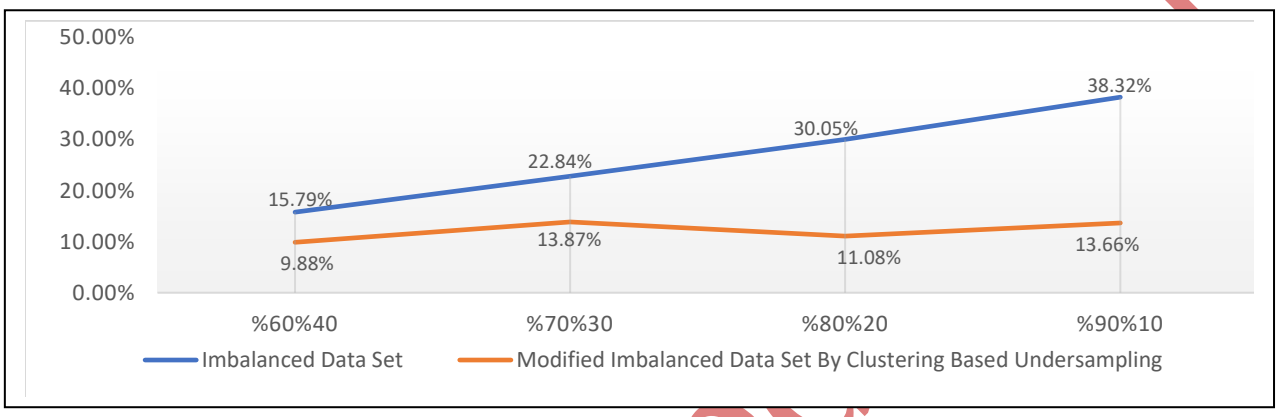

Fig 10. Comparison of type I error of SVM based on modified imbalanced and imbalanced data

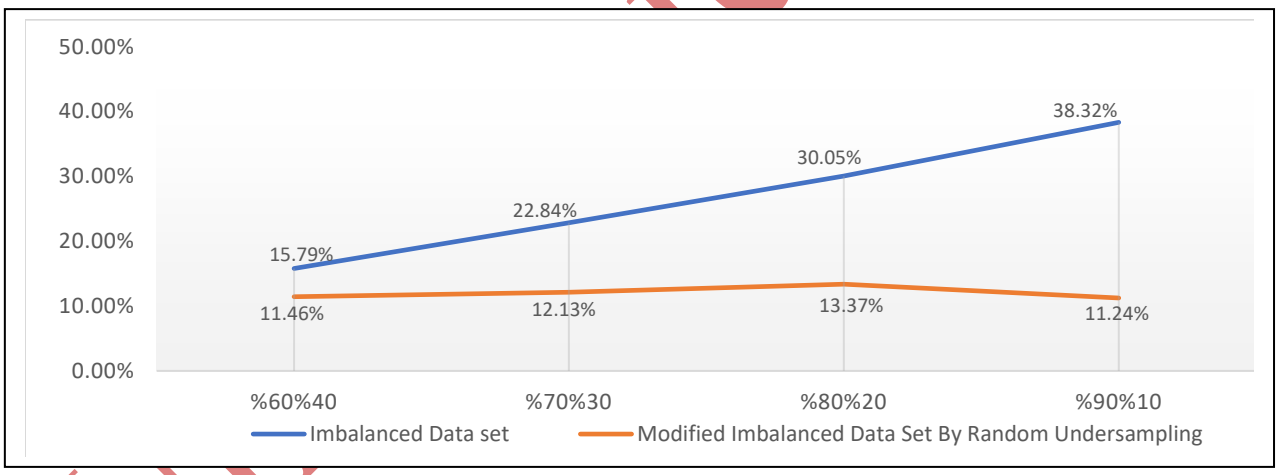

Fig 11. Comparison of type I error of SVM based on modified imbalanced and imbalanced data

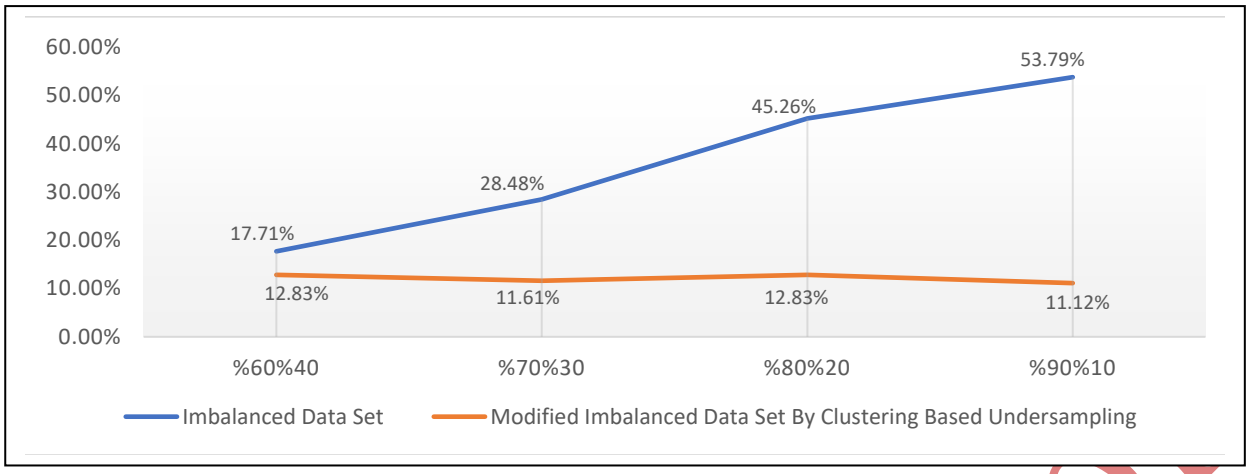

Fig 12. Comparison of type I error of Random Forest based on modified and imbalanced data

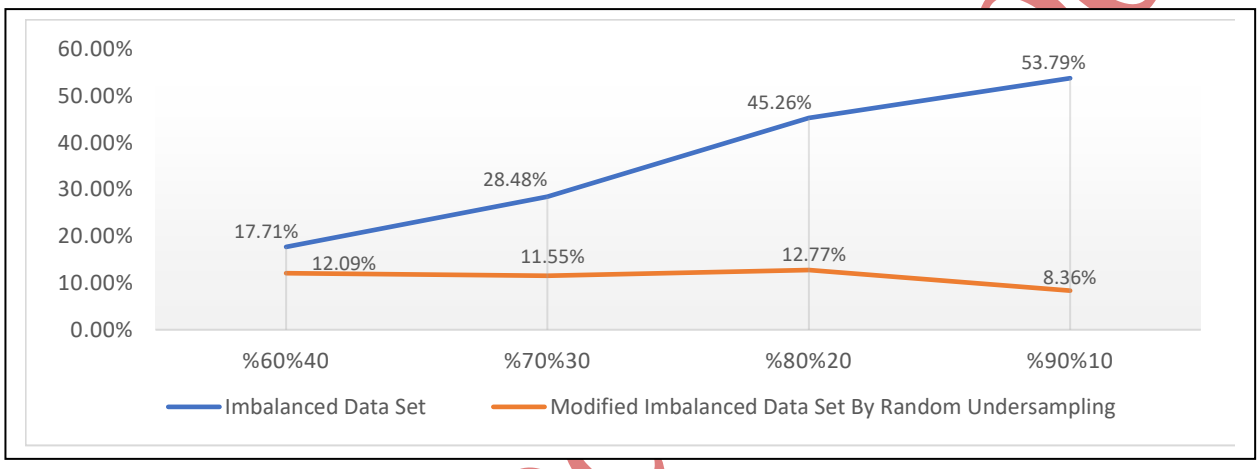

Fig 13. Comparison of type I error of Random Forest based on modified imbalanced and imbalanced data

H5 test result: According to Table 8, type II error (misidentification of a healthy company as a financially distressed company) of financial distress prediction model based on modified imbalanced data was higher compared to the models based on imbalanced data. As seen in Figure 14-17, as the data become more imbalanced , the blue line of the charts gained a downward trend. In other words, as the data became more imbalanced , type II error of financial distress prediction models based on imbalanced data got downward.

According to Table 8 and Figure 14 and 15, type II error of the support vector machine model based on modified imbalanced data by clustering-based undersampling method was significantly higher in comparison with imbalanced data with 4 different degrees at confidence levels 95% and 90%. Also, type II error of support vector machine model based on imbalanced data modified by random undersampling method was significantly higher compared to imbalanced data with degrees of 80% to 20% and 90% to 10% at 95% and 90% confidence levels. Therefore, applying the clustering-based method on training imbalanced data with 4 different degrees led to a significant increase in support vector machine model at 90% and 95% confidence levels and applying a random method on imbalanced training data with 80% degrees to 20% and 90% to 10% led to a significant increase in type II error of support vector machine model at 90% and 95% confidence levels.

According to Table 8 and figures 16 and 17, type II error of random forest model based on modified imbalanced data by clustering-based undersampling method was significantly higher compared to imbalanced data with 4 different degrees in 90% and 95% confidence levels. Also, type II error of random forest model based on modified imbalanced data by random undersampling method was significantly higher compared to imbalanced data with degrees of 70% to 30%, 80% to 20% and 90% to 10% at 95% and 90% confidence levels; so, applying the clustering-based method on imbalanced training data with 4

different degrees led to a significant increase in type II error of random forest model at 90% and 95% confidence levels. Applying random method on imbalanced training data with degrees of 70% to 30%, 80% to 20% and 90% to 10% led to a significant increase in type II error of random forest model at 90% and 95% confidence levels.

Table 8. Comparison of type II error based on balanced data models compared to modified imbalanced data

Financial distress prediction model	Ratio of imbalanced data	Type II error of modified imbalanced data		Type II error of imbalanced data	P-value (clustering method)	P-value (random method)
		random	clustering			
SVM	%60 VS %40	%9.2	%10.86	%6.56	**0.022	0.463
SVM	%70 VS %30	%8.86	%10.87	%6.83	**0.036	0.346
SVM	%80 VS %20	%12.06	%12.78	%4.10	**0.021	0.008**
SVM	%90 VS %10	%12.18	%11.27	%1.6	**0.008	0.012**
Random forest	%60 VS %40	%7.91	%9.49	%4.39	**0.008	0.144
Random forest	%70 VS %30	%9.61	%7.94	%1.88	**0.008	0.008**
Random forest	%80 VS %20	%9.16	%9.49	%0.33	**0.011	0.011**
Random forest	%90 VS %10	%8.92	%7.47	%0.044	**0.010	0.010**

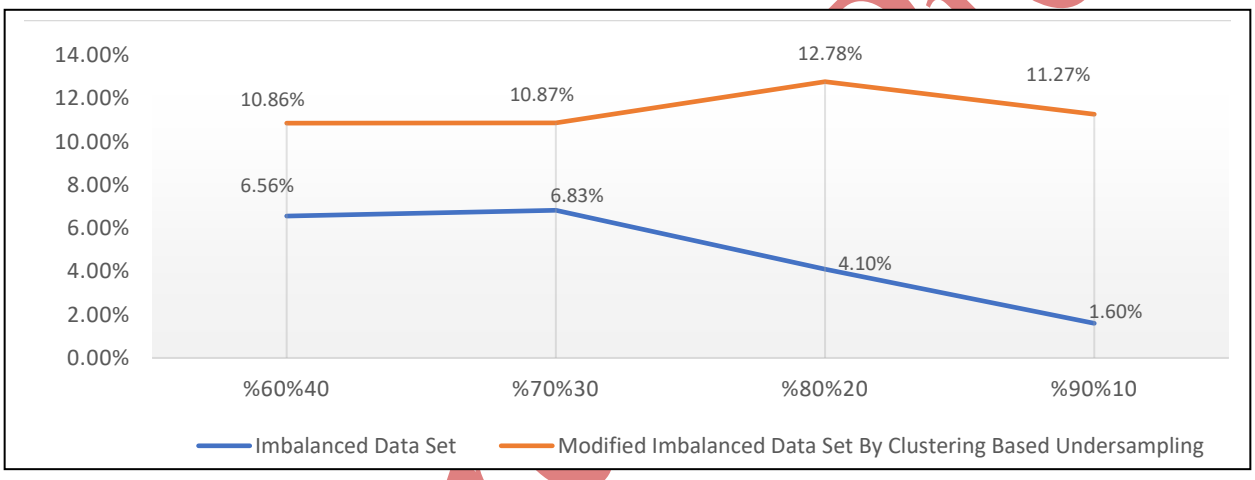

Fig 14. Comparison of type II error of SVM based on modified and imbalanced data

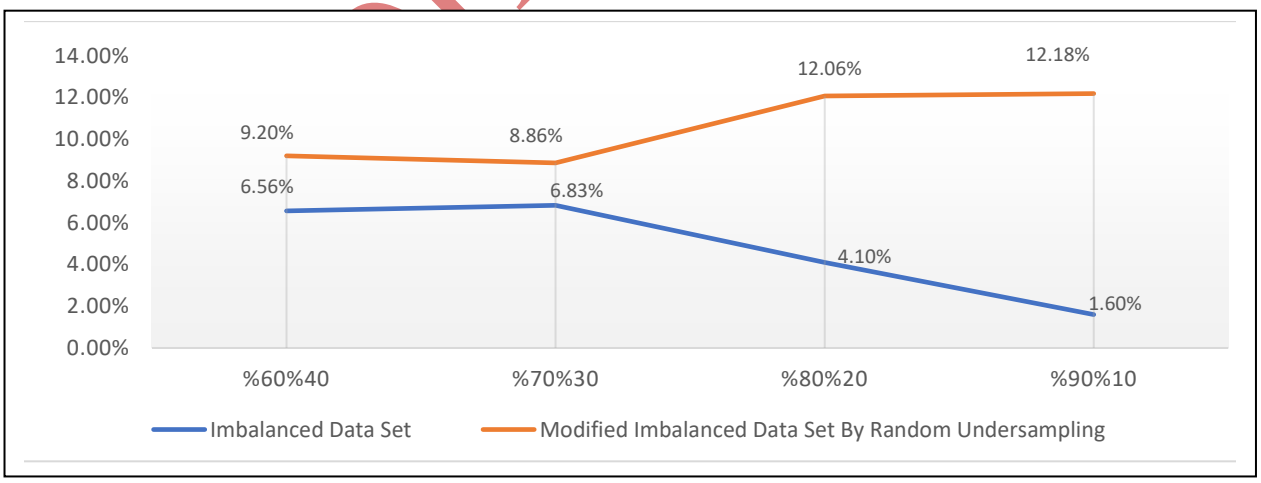

Fig 15. Comparison of type II error of SVM based on modified and imbalanced data

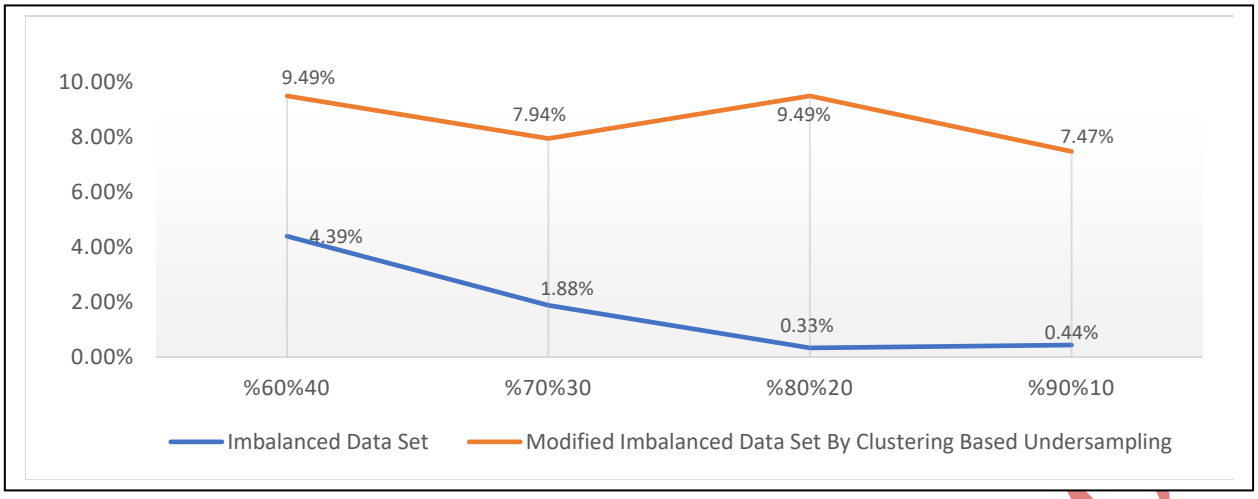

Fig 16. Comparison of type II error of Random Forest based on modified imbalanced and imbalanced data

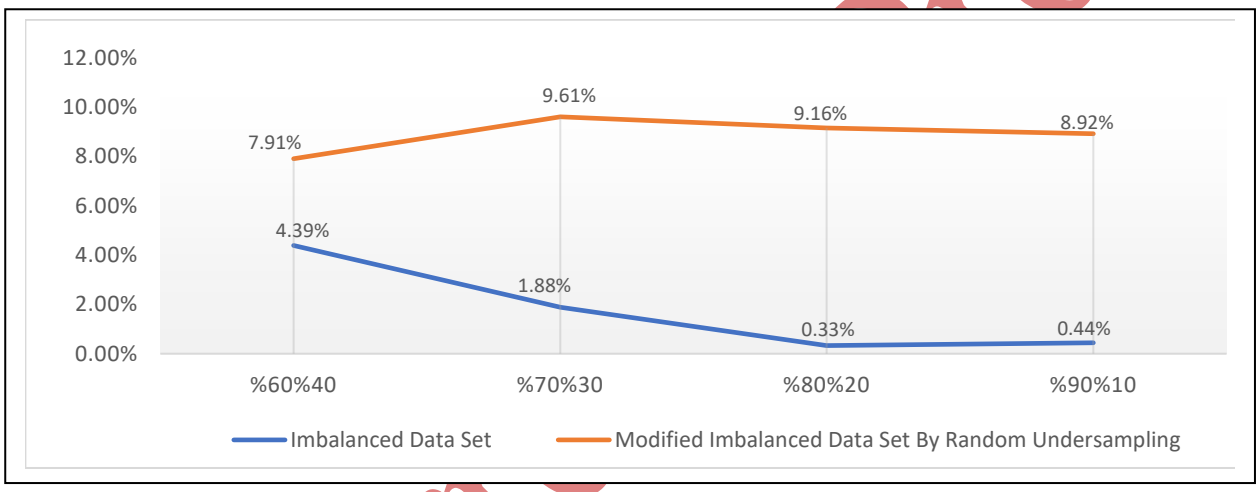

Fig 17. Comparison of type II error of Random Forest based on modified imbalanced and imbalanced data

H6 test result: According to Table 9, the geometric mean of financial distress prediction models based on modified imbalanced data was higher than the models based on imbalanced data. As seen in Figures 18 to 21, as the data become more imbalanced , the blue line of the graphs gets a downward trend. In other words, as the data becomes more imbalanced, the geometric mean of financial distress prediction models based on imbalanced data gets a downward trend.

According to Tables 9 and Figure 18 and 19, the geometric mean of the support vector machine model based on modified imbalanced data by clustering-based undersampling method was higher compared to imbalanced data with 4 different degrees; but, this advantage was not significant in 90% and 95% confidence levels; also, the geometric mean of the support vector machine model based on modified imbalanced data by random undersampling method was higher compared to imbalanced data; but, this advantage for imbalanced data with degrees 70% to 30% and 90% to 10% was significant at 90% and 95% confidence levels. Therefore, applying the random method to imbalanced training data with degrees 70% to 30% and 90% to 10% led to a significant increase in the geometric mean of the support vector machine model at 90% and 95% confidence levels.

According to Tables 9 and Figure 20 and 21, the geometric mean of the random forest model based on modified imbalanced data by clustering-based undersampling method and random undersampling method was higher than imbalanced data with 4 different degrees and this superiority is important for imbalanced data with degrees of 70% to 30%, 80% to 20% and 90% to 10% at 90% and 95% confidence levels; so, applying the clustering-based method and random method to imbalanced training data with 3 different degrees of 70% to

30%, 80% to 20% and 90% to 10% increased the geometric mean of the random forest model at 90% and 95% confidence levels.

Table 9. Comparing type geometric mean of models based on balanced data with modified imbalanced data

Financial distress prediction model	Ratio of imbalanced data	Geometric mean of modified imbalanced data		Geometric mean of imbalanced data	P-value (clustering method)	P-value (random method)
		random	clustering			
SVM	%60 VS %40	%89.59	%89.62	%88.67	0.531	0.835
SVM	%70 VS %30	%89.47	%87.61	%84.66	0.209	0.022**
SVM	%80 VS %20	%87.13	%87.98	%81.51	0.144	0.144
SVM	%90 VS %10	%88.18	%87.42	%76.26	0.144	0.037**
Random forest	%60 VS %40	%89.85	%90.30	88.68%	0.676	0.835
Random forest	%70 VS %30	%89.39	%90.19	%83.73	0.008**	0.008**
Random forest	%80 VS %20	%88.97	%88.78	%73.57	0.008**	0.008**
Random forest	%90 VS %10	%91.27	%90.61	%67.12	0.008**	0.008**

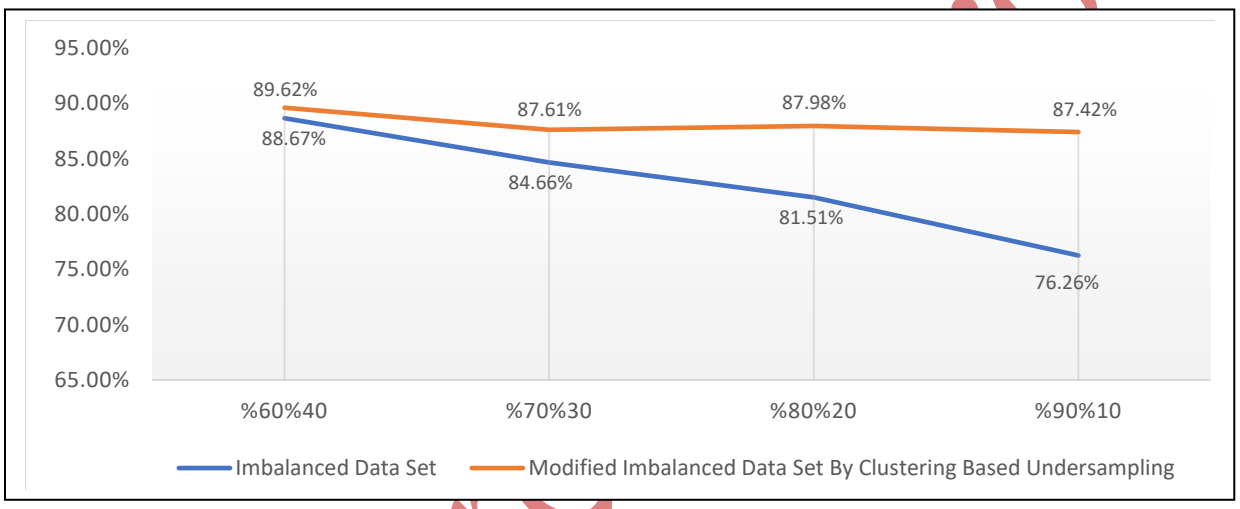

Fig 18. Comparison of geometric mean of SVM based on modified and imbalanced data

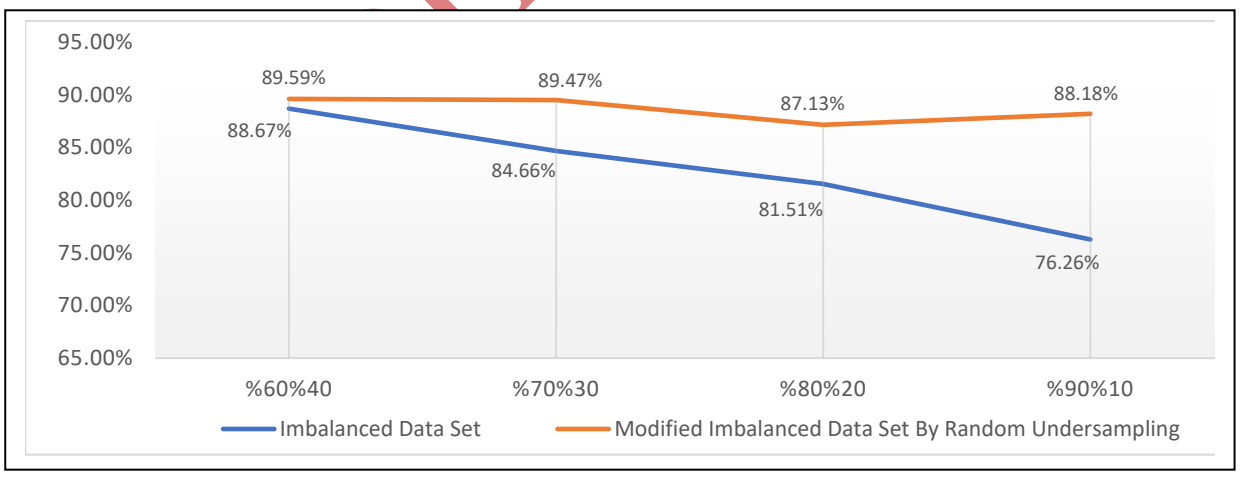

Fig 19. Comparison of Geometric Mean of SVM based on modified and imbalanced data

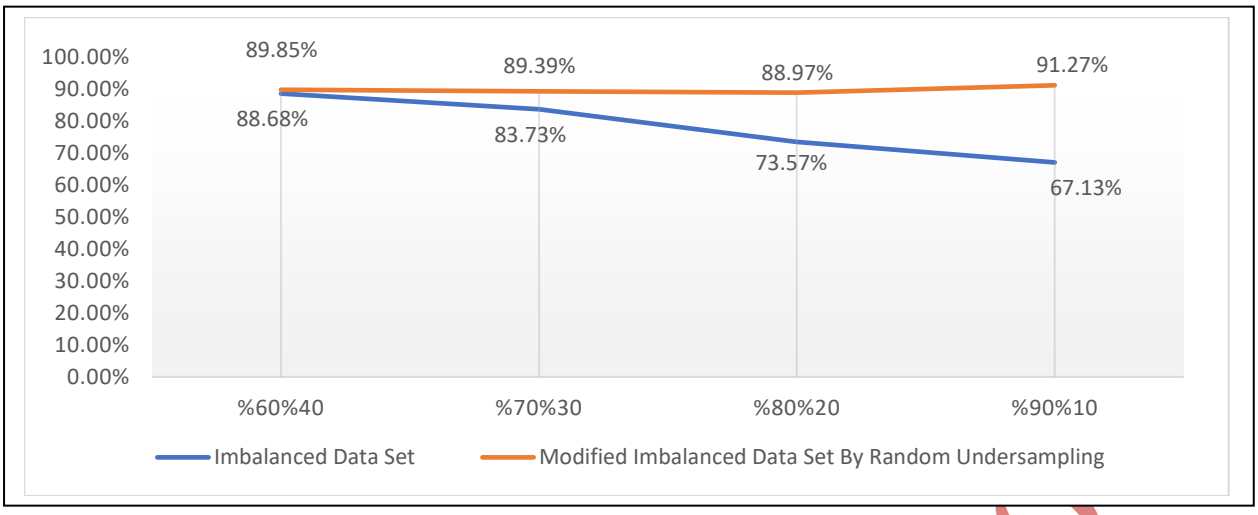

Fig 20. Comparison of geometric mean in Random Forest based on modified and imbalanced data

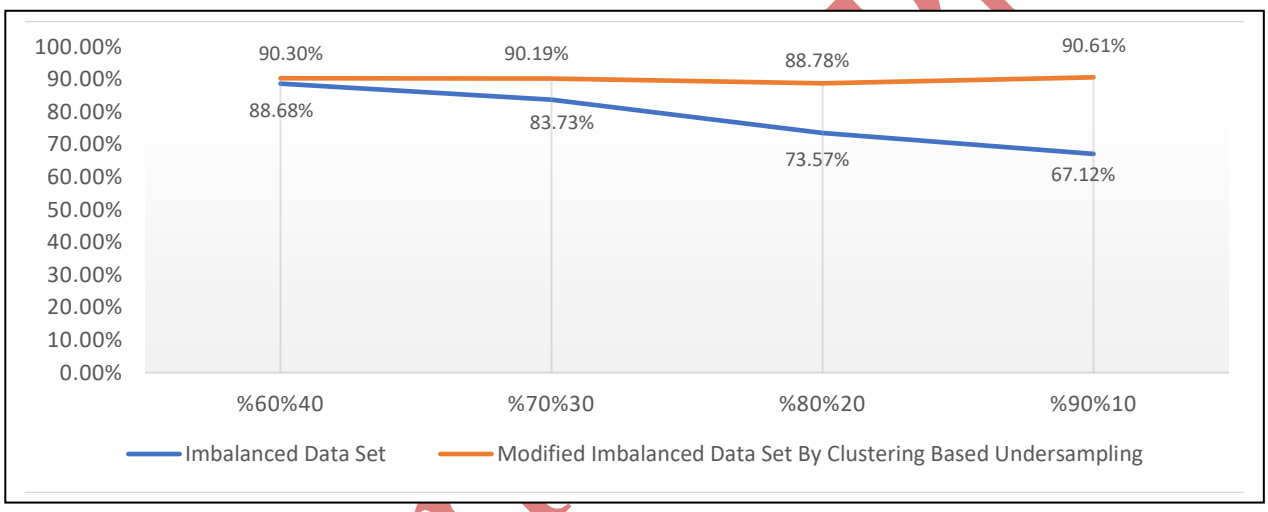

Fig 21. Comparing geometric mean in Random Forest based on modified and imbalanced data

4. Conclusion

H1 test results indicated that in all cases, type I error of financial distress prediction models based on balanced data is lower compared toimbalanceddata with 4 different degrees and in most cases, models 6 and 5 out of 8 financial distress prediction models at 90% and 95% confidence levels, respectively, type I error of balanced data-based models was less significant than imbalanced data at 90% and 95% confidence levels. H1 test result is consistent with those of Zmijewski [42] and Veganzones and Severin [40]. Zmijewski [42] showed that there is a negative relationship between the number (ratio) of financially distressed companies and the type I error. H2 test results indicated that in all cases, type II error of financial distress prediction models based on balanced data is more than imbalanced data with 4 different degrees and in most cases (models 8 and 7 out of 8 models of financial distress prediction at 90% and 95% confidence levels, respectively), type II error of financial distress prediction models based on balanced data is more significant than imbalanced data at 90% and 95% confidence levels. H2 test results are consistent with the results of Zmijewski [42] and Veganzones and Severin [40]. Zmijewski [42] showed a positive relationship between the number or proportion of financially distressed companies and type II error.H3 test results indicated that in most cases, the geometric mean of financial distress prediction models based on balanced data was higher than imbalanced data, and in models 4 and 3 out of 8 financial distress prediction models, respectively, the geometric Mean criterion of

models based on balanced data was significantly higher in comparision with imbalanced data at at 90% and 95% confidence levels at 90% and 95% confidence levels. These results consist with Veganzones and Severin [40] and Brown and Mues [7].

In interpreting H1to H3, it can be stated that lower type I error and higher type II error and the overall accuracy criterion of financial distress prediction models based on balanced data compared to imbalanced data with 4 degrees of 60% 40%, 70% to 30%, 80% to 20%, and 90% to 10% originate from the design of financial distress prediction models because financial distress prediction models are designed in such a way that the overall accuracy of the model is maximal. For this reason, in the training phase, the inclination is to the data of the majority class (healthy companies) and not much attention is paid to the data of the minority class (financially distressed companies) because the weight of the majority class (healthy companies) is higher and with maxizing the accuracy of the majority class (healthy companies) the overall accuracy of financial distress prediction models will be maximal[21]. In this study, the negative relationship between the number or ratio of financially distressed companies with the type of error and the positive relationship between the number or ratio of financially distressed companies with the type II error was confirmed[42]. In other words, the test results of the H1-H3 showed that with increasing data imbalance (decreasing the number of financially distressed companies in the selected sample), type I error has an upward trend and type II error has a downward trend, since the type I error and the type II error in financial distress prediction models based on balanced data compared to imbalanced data is lower and more than real, respectively; as mentioned, with the increase of data imbalance, I error has an upward trend and the type II error has a downward trend, but does the overall accuracy of financial distress prediction models increase or decrease? The results of testing the H3 show that in most cases except for 2 geometric means, the overall accuracy of financial distress prediction models based on balanced data is higher compared to imbalanced data and with increasing data imbalance (decrease in number of financially distressed companies), the geometric mean of the overall accuracy is declining and the geometric mean of the overall accuracy of the financial distress prediction models based on balanced data is higher than reality, especially for imbalanced data with higher degrees. Therefore, the use of imbalanced data-based prediction models is recommended to investors, creditors and other stakeholders because the type I error (identifying a financially distressed company as a healthy company) of the financial distress prediction model based on balanced data may be small, but the type I error of the same financial distress prediction model based on imbalanced data (reality) is high.

H4 test results showed that type I error of the support vector machine based on modified imbalanced data by clustering-based and random method is lower compared to imbalanced data which in most cases (models 6 and 4 of 8 financial distress prediction models respectively), are significantly lower at 90% and 95% confidence levels. Also, type I error of random forest model based on modified imbalanced data by clustering-based and random methods was lower compared to imbalanced data, which in most cases (models 6 and 7 out of 8 financial distress prediction models are significantly lower in 90% and 95% confidence levels. This result is consistent with the results of Veganzones and Severin [40]. H5 test results indicated that type II error of the support vector machine model based on the modified imbalanced data by clustering-based and random method was more than theimbalanceddata which in most cases (6 models of 8 support vector machine models), it was significantly higher at 90% and 95% confidence levels. Also, type II error of random forest model based on imbalanced data modified by clustering-based and random methods was more than imbalanced data which in most cases (7 and 6 models out of 8 random forest models) was significantly higher. This result is consistent with those of Veganzones and Severin [40]. H6 test results indicated that in all cases, geometric mean of the support vector machine model based on imbalanced data modified by clustering-based and random method was higher than imbalanced data and in some cases (6 models out of 8 financial distress prediction models), it was significantly higher at 90% and 90% confidence levels. This result is consistent with the results of Veganzones and Severin [40].

In interpreting H4 to H6, it can be stated that according to the result of Zmijewski [42], if the ratio or number of healthy companies to the ratio or number of financially distressed companies does not correspond to reality

or the data are balanced, sample selection bias may lead to underestimation of type I error and overestimation of the type II error of financial distress prediction models. Financial distress prediction models based on imbalanced data compared to balanced data will face the problem of higher type I error error (identifying a financially distressed company as a healthy company). The cost of the type I error is higher for investors, creditors and other stakeholders than the cost of the type II error of the financial distress prediction models. Therefore, type I error of financial distress prediction models based on imbalanced data should be reduced. In the present study, in order to reduce type I error (solving the problem of imbalanced data in predicting financial distress), random undersampling and clustering-based sampling have been used. Those sampling methods will lead to a reduction in the number (ratio) of healthy companies in training data (balancing training data), to financial distress prediction models in the training phase (learning model) to pay equal attention to both classes of healthy companies (majority class) and financially distressed companies (minority class). Therefore, H4 to H6 were formulated and tested. Their results indicated that by modifying the training process and balancing the data in the learning phase of the models (reducing the number of healthy companies in the training data) type I error, type II error and overall accuracy (by geometric mean) of the financial distress prediction models based on modified imbalanced data compared to the models based on imbalanced data have become lower, more and more, respectively based on expectations. Therefore, the use of financial distress prediction models based on modified unbalanced data is recommended; because the type I error, type II error and the geometric mean criterion of the overall accuracy of financial distress prediction models based on modified imbalanced data are based on reality. Then, type I error based on modified imbalanced data models is lower than imbalanced data; so, lower cost is imposed on investors, creditors and other stakeholders.

6. Practical Suggestion

1. Since the ratio of healthy companies to financially distressed companies (data imbalance) varies over time and the type of industry varies, and because performance (performance evaluation criteria) models of imbalanced financial distress prediction models depend on the degrees of data imbalance, investors and creditors are advised to choose financial distress prediction models according to the degree of data imbalance in the relevant time period or industry

2. Although the type I error and type II error of prediction models based on imbalanced data are suitable for real conditions, type I error of financial distress prediction models based on imbalanced data is higher than the models based on balanced data. The type I error imposes more costs on investors, creditors and other stakeholders compared to the second type of error; so, it is more important. Using undersampling method of training data, the learning process of imbalanced data-based models can be corrected and as a result, type I error and the geometric mean decreased and increased the overall accuracy of the financial distress prediction models, respectively; so, the use of financial distress prediction models based on the modified imbalanced data is recommended to investors, banks, credit financial institutions and other stakeholders because from one hand, type I error and type II error of the financial distress prediction models are compatible with reality because the test data of the models are imbalanced and on the other hand, the training data have been balanced and learning process of the model has been modified therefore the type I error has been reduced. The reduction of type I error will lead to a reduction in the cost of doubtful receivables of banks and credit financial institutions and a reduction in the cost will induce devaluation of investment.

7. Suggestions for Further Studies

1. In order to reduce type I error (misidentifying the financially distressed company as a healthy company) in this study, random undersampling and clustering-based sampling methods were used. For reducing type I error, oversampling and hybridization are suggested to be used. The use of sampling methods can reduce type I error and increase overall accuracy, leading to optimal decisions for investors, creditors, and other stakeholders.

2. The issue of imbalanced data in financial distress prediction is due to the unequal distribution of healthy and financially distressed companies. It is suggested that in future research, performance evaluation criteria of fraud detection models be compared based on balanced data, imbalanced data, and modified imbalanced data;

also, in order to reduce type I error (misidentification of a fraudulent company as a non-fraudulent company) oversampling, undersampling, and combined methods are suggested to be used.

3. As mentioned earlier, type I error and type II error of financial distress prediction models based on imbalanced data are more and less, respectively compared to balanced data-based models. The type I error of the financial distress prediction models has a higher cost for investors and creditors compared to type II error. In the present study, for reducing type I error, undersampling methods of training data models were used. Further studies are suggested to use cost-sensitive learning method to reduce the type I error and increase the geometric mean criterion.

References

[1] Altman, E, I., *Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy*, Journal of Finance, 1968. 23(4),P. 589-609. Doi: 10.2307/2978933.

[2] Anderson, R., *The credit scoring toolkit: Theory and practice for retail credit risk management and decision automation*, Oxford University Press, 2007.

[3] Anwar, M. N., *Complexity measurement for dealing with class imbalance problems in classification modelling*, Thesis for Doctor of Philosophy, Massey University, Institute of Fundamental Sciences, 2012.

[4] Balcaen, S., Ooghe, H., *35 years of studies on business failure: an overview of the classic statistical methodologies and their related problems*, British Accounting Review, 2006, 38(1), P. 63-93. Doi: 10.1016/j.bar.2005.09.001.

[5] Beaver, W., *Financial Ratios as Predictor of Failure*. Journal of Accounting Research, 1966. 4, 71-111.

[6] Breiman, L., Random Forests. Machine Learning, 45(1), P. 5-32, 2001.

[7] Brown, I., Mues, C., *An experimental comparison of classification algorithms for imbalanced credit scoring data sets*. Expert Systems with Applications, 2012, 39(3), P.3446-3453 .Doi: 10.1016/j.eswa.2011.09.033

[8] Buda, M., *A Systematic Study of the Class Imbalance Problem in Convolutional Neura Networks*. Royal Institute of Technology, School of Computer Science and Communication, Sweden, 2017.

[9] Chawla, N. V., Japkowicz, N., Kotcz, A., Editorial: *Special issue on learning from imbalanced data sets*. ACM Sigkdd Explorations Newsletter., 2004, 6(1), P. 1–6. Doi: 10.1145/1007730.1007733.

[10] Chawlaet, N. V., (2005). *Data mining for imbalanced datasets*: *An overview*, Data Mining and Knowledge Discovery Handbook, 2009, P. 875-886, Doi: 10.1007/978-0-387-09823-4_45.

[11] Chen, H.-J., Huang, S. Y., Lin, C.-S., *Alternative diagnosis of corporate bankruptcy*: A neuro fuzzy approach. Expert Systems with Applications, 2009, 36(4), P.7710-7720, Doi: Doi.org/10.1016/j.eswa.2008.09.023

[12] Faris, Hossam., Abukhurma, Ruba., Waref, Almanasee,. Saadeh, Mohammed., Mora, Antonio M., Castillo, Pedro A., Aljarah, Ibrahim., *Improving financial bankruptcy prediction in a highly imbalanced class distribution using oversampling and ensemble learning: a case fromthe Spanish market.* Artificial Intelligence, 2020(9), P.31-53. Doi: 0.1007/s13748-019-00197-9.

[13] García, S., Herrera, F., *Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy*. Evolutionary Computation, 2009, 17(3), P.275–306. Doi: doi.org/10.1162/evco.2009.17.3.275.

[14] Ghasemi, S., Sarlak, A., *Investigating the Impact of the Financial Crisis on Conservative Accounting and Transparency of Banking Information*. Advances in Mathematical Finance & Applications, *3(3)*, 2018, P.53-68 (in Persian). Doi: 10.22034/AMFA.2018.544949.

[15] Ghatasheh, N., Hossam, F., Abukhurma, R., Castillo, P., Al-Madi, N., Mora, A., Hassanat., *Cost-sensitive ensemble methods for bankruptcy prediction in a highly imbalanced data distribution: a real case from the Spanish market*. Progress in Artificial Intelligence, 2020, 9, P.361-375. Doi: 10.1007/s13748-020-00219-x.

[16] Gordini, N., *A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy*. Expert Systems with Applications, 2014, 41(14), P.6433-6445, Doi:10.1016/j.eswa.2014.04.026

[17] Haghparast, A., Momeni, A., and Gerd, A., *Visual financial ratios and bankruptcy prediction of companies using convolutional neural network model*, financial engineering and Tehran Stock Exchange, 2021, 12 (46), P.558-575(in Persian), DOR: 20.1001.1.22519165.1400.12.46.24.0

[18] Heidari Farahani, M., Ghayur, Farzad., and Mansourfar, Gh., The effect of management behavioral aspects on financial distress. Financial accounting research. 2020, 11(3) ,P.117-134 (in Persian), Doi:1 0.22108/far.2020.119602.1534

[19] Hsu, C.W., Chang, C.C., Lin, C.J., *A Practical Guide to Support Vector Classification. Technical Report*, Department of Computer Science and Information Engineering, National Taiwan University, 2004.

[20] Khoshtinat, M., and Qasouri, M., *Comparison between hybrid financial ratios based on cash flows and accruals with financial ratios based solely on accruals in predicting corporate bankruptcy*. Empirical Studies in Financial Accounting, 2005, 9 (3), P.43-61.

[21] Kim, M. J., Han, I., *The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms*, Expert Systems with Applications, 2003, 25(4), P.637-646. Doi: 10.1016/S0957-4174(03)00102-7

[22] Kim, T., Ahn, H., *A hybrid undersampling approach for better bankruptcy prediction*. Journal of Intelligence and Information Systems, 2015, 21(2), P.173-190.

[23] Kotsiantis S., Pintelas P., *Mixture of expert agents for handling imbalanced data sets*, Annals of Mathematics, Computing & TeleInformatics, 2003, 1(1) ,P.46–55.`

[24] Li, H., Sun, J., Ranking-order case-based reasoning for financial distress prediction. Knowledge-based Systems, 2008, 21(8), P.868–878. Doi: 10.1016/j.knosys.2008.03.047

[25] Lin, SW., Ying, KC., Chen, SC., Lee, ZJ., Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Systems with Applications, 2008, 35. P.605-617, Doi: 10.1007/978-3-319-13563-2_51

[26] Lopez, V., Fernández, A., García, S., Palade, V., Herrera, F., *An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics*. Information Sciences, 2013, 250, P.113-141.Doi: 10.1002/(SICI)1099-131X(200004)19:3<219::AID-FOR752>3.0.CO;2-J

[27] Mann, H. B., Whitney, D. R., *On a test of whether one of 2 random variables is stochastically larger than the other*. Annals of Mathematical Statistics, 1947, 18(1), P.50-60

[28] Mansourfar, Gh., Ghayur, F., Lotfi, B., *The ability of support vector machine to predict financial distress*, Empirical Accounting Research, 2015, (17) 5, P.177-195 (in Persian).

[29] McKee, T. E., Greenstein, M., *Predicting Bankruptcy Using Recursive Partitioning and a Realistically Proportioned Data Set*. Journal of Prediction, 2000, 19(3), 219-230. Doi: 10.1002/(SICI)1099-131X(200004)19:3<219::AID-FOR752>3.0.CO;2-J

[30] Mohseni, R., Agha Babaei, R., and Ghorbani, V.M., *Financial distress prediction with using efficiency as a predictor variable*, Quarterly Journal of Economic Research and Policy, 1392, 21(65), P.14-123 (in Persian).

[31] Newton, G. w., *Bankruptcy and Insolvency Accounting, practic and procedure*. Volume1, John Wiley & Sons, Inc. Seventh Edition. 2010.

[32] Olson, D. L., Delen, D., Meng, Y., *Comparative analysis of data mining methods for bankruptcy Prediction*, 2012, P. 464.473.Doi: 10.1016/j.dss.2011.10.007.

[33] Ooghe, H., and P. Joos. *Failure prediction, explanation of misclassifications and incorporation of other relevant variables:* result of empirical research in Belgium. Working paper, Department of Corporate Finance, Ghent University (Belgium), 1990

[34] Raei, R., Fallahpour, S., *Financial distress prediction of companies using artificial neural network*. Journal of Financial Research, 2004, 6 (1), P.39-69 (in Persian).

[35] Razavi, B., Mehrazin, A, R., Shoorvarzi, M, R., Massihabadi, A., *Experimental Comparison of Financial Distress Prediction Models Using Imbalanced data sets*, Advances in Mathematical Finance in Applications 2022, 7(3), (in Persian), 10.22034/AMFA.2021.1905055.1461

[36] Rezaei, N., Javaheri, M., *The Predictability of Neural Network and Genetic Algorithm from Companies' Financial Crisis*, Advances in Mathematical Finance in Applications, 2020, *5(2)*, P.183-196 (in Persian). Doi: 10.22034/AMFA.2019.1863963.1195

[37] Saruei, S., *The Study of Performance of Springerit, Zimsky and Ahlson Models in Predicting Bankruptcy of Listed Companies in Tehran Stock Exchange*, M. A. thesis, Arak Islamic Azad University, Arak, Iran, 2010 (in Persian).

[38] Thabtah, F., Kamalov, F., Rajab, K., *A new computational intelligence approach to detect autistic features for autism screening*. International Journal of Medical Infromatics, 2018, P.112-117. Doi: 10.1016/j.ijmedinf.2018.06.009.

[39] Vapnik, V., Statistical Learning Theory (Vol. 2). New York: Springer, 1998.

[40] Veganzones, D., Severin, E., *An investigation of bankruptcy prediction in imbalanced datasets.* Decision Support System, 2018, 112, P.111-124. Doi: 10.1016/j.dss.2018.06.011

[41] Wei-Chao, L., Chih-Fong, Tsai., Ya-Han, Hu., Jing-Shang, Jhang., Clustering-based undersampling in class-imbalanced data,. Information Sciences, 2017, 409, P.17-26.

[42] Zmijewski, M. E., *Methodological issues related to the estimation of financial distress prediction models.* Journal of Accounting Research, 1984, 22, P.59–82. Doi: 10.2307/2490859

[43] Zoricák, M., Gnip, P., Drotár, P., Gazda, V., *Bankruptcy prediction for small and medium-sized companies using severely imbalanced datasets*. Economic Modelling, 2020, 84, P.165-176. Doi: 10.1016/j.econmod.2019.04.003.