INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Impact of Different Outlier Handling Techniques on GAN Based Hybrid Bankruptcy Prediction Models

Sasmita Manjari Nayak[1], **Minakhi Rout**[1]*

1 School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha, India

OPEN ACCESS

Received: 20-03-2023

Accepted: 14-12-2023

Published: 22-01-2024

Citation: Nayak SM, Rout M (2024) Impact of Different Outlier Handling Techniques on GAN Based Hybrid Bankruptcy Prediction Models. Indian Journal of Science and Technology 17(4): 373-385. https://doi.org/10.17485/IJST/v17i4.649

* **Corresponding author**.

minakhi.routfcs@kiit.ac.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: This study focuses on the crucial task of predicting bankruptcy within the framework of contemporary economics. It is essential to construct trustworthy and solid models in order to accomplish this. **Objective s:** The main goal is to address the intricacies involved in predicting bankruptcy by putting forth models that can manage scenarios with a high concentration of outliers and data imbalances. **Methods:** We use a Generative Adversarial Network (GAN) to synthesise data in order to address data imbalance. Then, we deal with outliers using six different methods: z-score normalisation, isolation forest, one-class SVM, local outlier factor, elliptical envelope, and interquartile range (IQR). Two different bankruptcy datasets are subjected to rigorous application of these suggested models, and their performance is carefully compared. **Findings:** Compared to the other three models, the GAN-Adaptive Neuro-Fuzzy Inference System (ANFIS) model outperforms the others in terms of bankruptcy prediction. This model is remarkably resilient to outlier data; regardless of the use of outlier handling approaches, its predicted accuracy does not change. **Novelty & Applications:** The innovation resides in the introduction of ANFIS models based on GANs, which are explored in the field of bankruptcy dataset prediction that is tainted by outliers. Furthermore, although a number of outlier handling strategies have been investigated in earlier research, our study is a ground-breaking attempt to pinpoint the best strategies for enhancing bankruptcy prediction with hybrid ANFIS models.

Keywords: Bankruptcy; Feature Selection (FS); GAN (Generative Adversarial Network); Adaptive NeuroFuzzy Inference System (ANFIS); Outlier handling (OH)

1 Introduction

Since effective bankruptcy forecasting is crucial to a company's future existence, bankruptcy prediction can be regarded as one of the major research areas. For corporations to take corrective action in a timely manner and prevent bankruptcy, they need accurate prediction models. In general, outliers can be found in bankruptcy statistics, and the majority of them exhibit imbalances. These factors make pre-

processing the datasets crucial before creating models to predict bankruptcy.

An outlier is a small amount of data that stands out significantly from the rest of the data. Outliers in a dataset behave differently from the rest of the data in that dataset. Outlier handling is therefore crucial in many different application domains. This study makes use of two datasets. In general, there are a lot of outliers in bankruptcy datasets. Here are six methods for addressing outliers: Using a comparison of one class SVM, local outlier factor, isolation forest, elliptical envelope, interquartile range (IQR), and z-score, the optimal outlier treatment method for bankruptcy datasets is determined. We use one data generating technique, GAN, to balance the datasets because they are also very imbalanced in nature. Here, a hybrid model for predicting bankruptcy named GAN-ANN, GAN-LSTM, GAN-CNN, and GAN-ANFIS is introduced. Applying all of the outlier handling techniques to them, we compare them in order to determine which classification model best predicts bankruptcy.

2 Literature Review

Outliers are typically present in bankruptcy datasets, which may have an impact on the prediction rate. An outlier is a small amount of data that stands out significantly from the rest of the data. Outliers in a dataset behave differently from the rest of the data in that dataset. Therefore, after detection, outliers need to be treated appropriately. Local outlier factor (LOF) was employed by the authors in [1,2] as an outlier detection technique, and this resulted in a superior prediction. Z-score was successfully employed as an OH approach in [3] by researchers using Multivariate Long Short Term Memory (MLSTM). The authors handle outliers using both the Z-score and the local outlier factor [4]. The authors of [5,6] employed the interquartile range approach (IQR) to address an outlier and discovered that the performance of the categorization has improved after the outliers in the data were removed. Isolation Forest was employed by the authors in [7] as an outlier handling method. In [8], the Elliptic Envelope approach is employed as an OH technique. In contrast, the authors applied two unsupervised machine learning techniques—the Elliptic Envelope (EE) and the Isolation Forest (If)—to improve the prediction results [9,10]. One-class SVM was the only method employed by the author in [11] for OH; however, in [12], three algorithms—one-class SVM, elliptic envelope, and local outlier factor—were chosen and combined based on a vote process. We discovered that the authors of the aforementioned studies use various outlier handling strategies to improve their predicted outcomes. Here, we analyse six outlier treatment methods to determine which is best: elliptical envelope, one class SVM, z-score, isolation forest, local outlier factor, and interquartile range (IQR). A comprehensive analysis is being carried out in this study to ascertain the effects of utilising various OH techniques and to identify which one is more suited for bankruptcy datasets.

For the purpose of predicting bankruptcy, the authors of [13] compare three machine learning techniques—Random Forest, SVM, and Logistic Regression—and one deep learning technique, CNN. It is discovered that when it comes to bankruptcy prediction, deep learning algorithms perform better than conventional models. The identical thing is also present in [14]. The author discovered that deep learning techniques, such as RNN and LSTM, have greater predictive power than all conventional machine learning models after comparing them to a few other machine learning techniques. Using the adaptive neuro-fuzzy inference system (ANFIS) as a classifier, the researchers in [15] compared the ANFIS's performance to that of the genetic algorithm as well as other statistical techniques like support vector machines, Bayesian networks, and J48 decisions. They discovered that the ANFIS produced better results than the other techniques. Research on the measuring and management of banking risk is conducted in [16], wherein the researchers achieved excellent results using ANFIS as a classifier. The researchers in [17,18] achieved higher prediction accuracy by using ANFIS as a classifier for bankruptcy prediction. Some deep learning researchers have shown in the aforementioned studies that deep learning techniques outperform machine learning techniques in the context of bankruptcy. Additionally, some scholars are working on the ANFIS model. The ANFIS model consistently provides very accurate predictions. However, no one is attempting to determine whether of the deep learning and ANFIS approaches has greater predictive potential. In order to determine the best classification model for bankruptcy prediction, we introduce four hybrid models for this region: GAN-ANN, GAN-LSTM, GAN-CNN, and GAN-ANFIS. As a balancing technique, we use the data generating technique GAN (Generative Adversarial Network) [19,20]. Therefore, the purpose of this study is to determine the most effective outlier handling method and classification model for bankruptcy datasets.

3 Methodology

In general, outliers can be found in bankruptcy statistics, and the majority of them exhibit imbalances. In this study, six outlier handling techniques—one class SVM, local outlier factor, isolation forest, elliptical envelope, interquartile range (IQR), and z-score—were used for the pre-processing of datasets. One data-generating approach, GAN, was also employed to balance the dataset. Since more inputs increase the complexity of the model, we are only using two characteristics from the datasets for the ANFIS model. Here, we choose features based on feature relevance in order to select the most relevant features.

3.1 Local outlier factor (LOF)

The local outlier factor (LOF) of each object is determined by assigning a degree to it in the LOF technique. This method, which is mostly focused on local density, finds outliers by comparing the object's local density with that of its closest neighbour. A low-density neighbourhood is indicated by a high LOF number, which increases the likelihood of an outlier [2].

3.2 Z-score

z-score measures the deviation of different experimental values from the most probable value, the mean. Z- score can be calculated by applying the following formula.

$Z = (i- \mu)/\sigma$

where,

i= Input value

μ = Mean of the data set

σ = Standard Deviation of the dataset

We may determine how much one observation deviates from the other observations in the dataset using the z-score value; if the value is large, the observation will be considered an outlier. This method is quite straightforward and quick to apply, but it requires some understanding of how the data set is distributed [4].

3.3 Interquartile Range (IQR)

Outliers can only be found using the interquartile range (IQR) on uniformly distributed data. The 25th, 50th, and 75th values in an IQR dataset are the four groups into which the dataset is divided, or into four quartiles. For the computation of outliers First, the data must be maintained in the proper order using the Interquartile Range. Next, by determining the lower 25% and upper 75% of the distribution, the values of the first quartile (Q1) and third quartile (Q3) are determined. Next, using the following formula, the values of the upper limit (UL), lower limit (LL), and IQR are determined.

IQR = Q3-Q1

LL = Q1-(1.5*IQR)

UL = Q3+(1.5*IQR)

Everything outside the upper and lower boundary is an outlier [6].

3.4 Isolation Forest

In an isolation forest, data is divided at each node using a randomly selected feature and threshold to create a tree structure that grows from the root to the leaves. Every tree develops until every occurrence is by itself in a leaf. In this case, the instances are split repeatedly; anomalous or out-of-the-ordinary data quickly arrive at leaf nodes, but nominal data require an increasing number of splits before they do. Therefore, there is a very high chance that a forest of casual trees or a forest of random trees that together provide shorter path lengths for some specific places or observations are outliers [7].

3.5 Elliptical envelope

The elliptic envelope method of outlier detection assumes that the normal data items constitute a distribution, such as the Gaussian distribution. It uses the data points' normal distribution along with any possible feature covariance. Based on the supposition that normal objects occur in high probability sections of the distribution and outliers occur in low probability regions or do not follow this distribution, the data objects are separated into normal data and outliers in this instance [8].

3.6 One class SVM

In support vector machines (SVM), nonlinear maps can be used to convert nonlinearly separable data into a high-dimensional space where the points can be maintained linearly apart. The SVM methodology can be applied to the one-class classification problem using the one-class SVM technique [11].

3.7 Adaptive Neuro-Fuzzy Inference System (ANFIS)

The fuzzy system and the artificial neural network model are combined to create the hybrid model known as ANFIS. The fuzzy layer, product layer, normalised layer, de-fuzzy layer, and total output layer are the five main layers of the ANFIS. In the first

layer, each node is an adaptable node. Membership functions like the Gaussian membership function and the generalised bell membership function are typically used here as node functions. The firing strength of a rule is represented by each node output in the second layer. Each rule's normalised firing strength is displayed on each layer 3 node. Each adaptive node in layer 4 has a node function that indicates how the rules affect the final product as a whole. A single node in Layer 5 calculates the sum of all rule outputs [17,18].

3.8 GAN (Generative Adversarial Network)

GAN, a generative technique, generates new data instances that resemble the training set. For example, we can create photos with GANs that don't have any faces that resemble those of actual humans. The discriminator and generator neural networks in a GAN compete with one another to become more accurate. The goal of a GAN is to teach a discriminator that must be able to distinguish between real and fake data while also giving instructions to a generator to produce fake data instances that can reliably fool the discriminator [19,20].

3.9 Feature selection on Feature Importance

It is the method by which we select the features—either manually or automatically—that contribute most to the target variable. It describes techniques that assign a number to input features according to how well they can predict target variables; feature selection is then done using the feature Importance score. The improvement of a predictive model's efficacy and efficiency is greatly influenced by feature significance scores [21]. The datasets used in this study are split into two categories after preprocessing: training and testing. Here, the models are trained using 80% of the entire data, with the remaining 20% being utilised to assess the models' efficacy. To train the model, the preprocessed training dataset is fed into the classification models.

3.10 Architecture

The process of doing the pre-processing of the datasets, construction of bankruptcy prediction models and their evaluation can be represented pictorially step by step as shown in Figure 1.

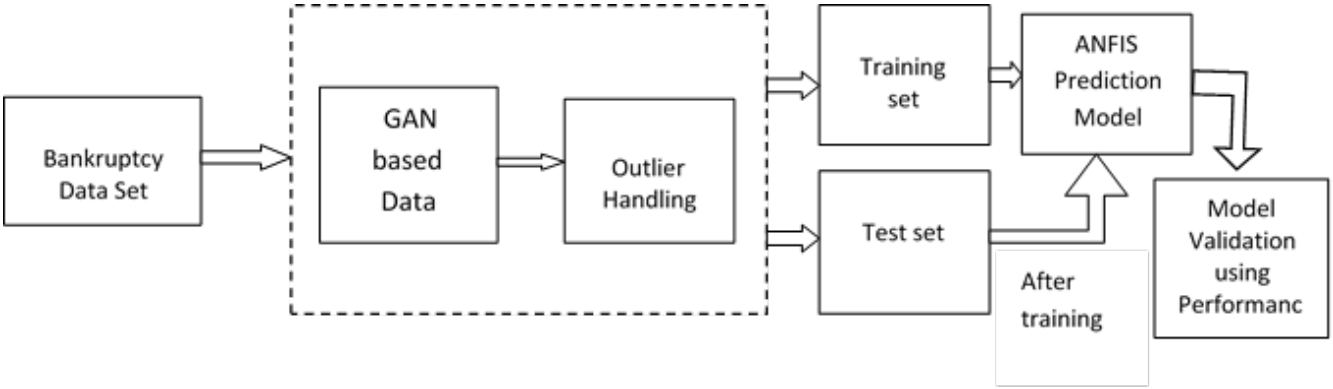


Fig 1. The architecture of GAN-ANFIS bankruptcy prediction process

4 Dataset Description

Two datasets are being used here: the first is for financial distress prediction, with 87 characteristics and 3682 occurrences, of which 3546 are nonbankrupt and 136 are bankrupt. In this case, the first column represents several sample companies, while the second column displays various time periods to which the data pertains. The duration of the time series varies from 1 to 14 for every organisation. Other features of the sampled companies include some financial and non-financial ones.

The second is the Taiwan Stock Exchange, which has 6816 cases with 96 attributes, 220 of which are bankrupt, and 6599 of which are not. The information used here was gathered from the Taiwan Economic Journal between 1999 and 2009. The definition of company bankruptcy was established by the business legislation of Taiwan.

5 Results and Discussion

The datasets used in this study are split into two categories: training and testing. Here, the models are trained using 80% of the entire data, with the remaining 20% being utilised to assess the models' efficacy. This study's datasets are incredibly unbalanced. The GAN (Generative Adversarial Network) data generating method is used to balance the unbalanced training datasets by producing the necessary number of minority class data. Subsequently, we employ an additional dataset preprocessing technique called outlier treatment to eliminate outliers from the datasets. The preprocessed training dataset is used to train the models in classification after preprocessing.

Here, four classification models are employed: ANN, LSTM, CNN, and ANFIS. We use Conv1D () in our CNN model. It so has a convolution layer that is one dimension. We apply 128 filters to the dataset in the first Conv1D () layer, with a convolutional window size of 3. The number of features to be chosen from the dataset for prediction is determined by the input_shape parameter. Lastly, a rectified linear activation function (ReLU) activation function is employed in four layers of hidden neurons. Here, two is used as the pool_size. The output then appears, showing us a dense layer with two neurons and a constant pair of projected values that must be either 0 or 1. Since the softmax activation function goes from 0 to 1, it is utilised in this situation and makes it simple to predict a binary value as the output. The data in this model is transformed into a one-dimensional array using the Flatten() function in order to move on to the next phase of processing. We have a successively dense LSTM model. Here, there are 80 neurons in the first LSTM layer. The number of features in the dataset that are utilised for prediction is determined by the input shape. Then another layer was added. Lastly, there are two neurons in the output layer since we require both bankrupt and nonbankrupt outputs. To prevent overfitting, we add a few Dropout layers after the LSTM layer. For the Dropout layers, we set a value of 0.2, meaning that 20% of the layers will be eliminated. Next, a single output unit is specified by the Dense layer. This uses the sigmoid activation function. We have two hidden layers, one output layer, and one input layer in our ANN model. Here, the number of neurons in the input layer fluctuates depending on how many characteristics are present in the dataset that is used for prediction. In contrast, the output layer always contains two neurons because we require both bankrupt and nonbankrupt outputs. In this case, there are 20 neurons in one buried layer and 10 in the other.

The same functions and parameters are used to compile each of the aforementioned three models. A loss function called "sparse_categorical_crossentropy" is used, and an Adam optimizer with a learning rate of 0.001 is employed initially to optimize the loss function. Adam is an adaptive learning rate method that calculates personal learning rates based on a range of factors. Adam uses estimates of the first and second moments of the gradient to modify the learning rate for each neural network weight, hence the name "adaptable moment estimation." We use a batch size of ten when running our models to train them. Additionally, 50 epochs are specified here.

There are five layers in our ANFIS. For every input variable, the Gaussian membership function was employed. Since we only require one output at a time, there is only one node in the fifth layer, the output layer. We use a two-pass learning technique in our ANFIS model. The first pass is a forward pass, where nodes' outputs are calculated up to the fourth layer, and subsequent parameters are updated using least squares techniques. Mistakes are carried backward in the second pass until they reach the first layer, at which point ANFIS applies gradient descent to optimise the membership function's parameters. We execute the model with the epoch number set to 50 in order to train our ANFIS model.

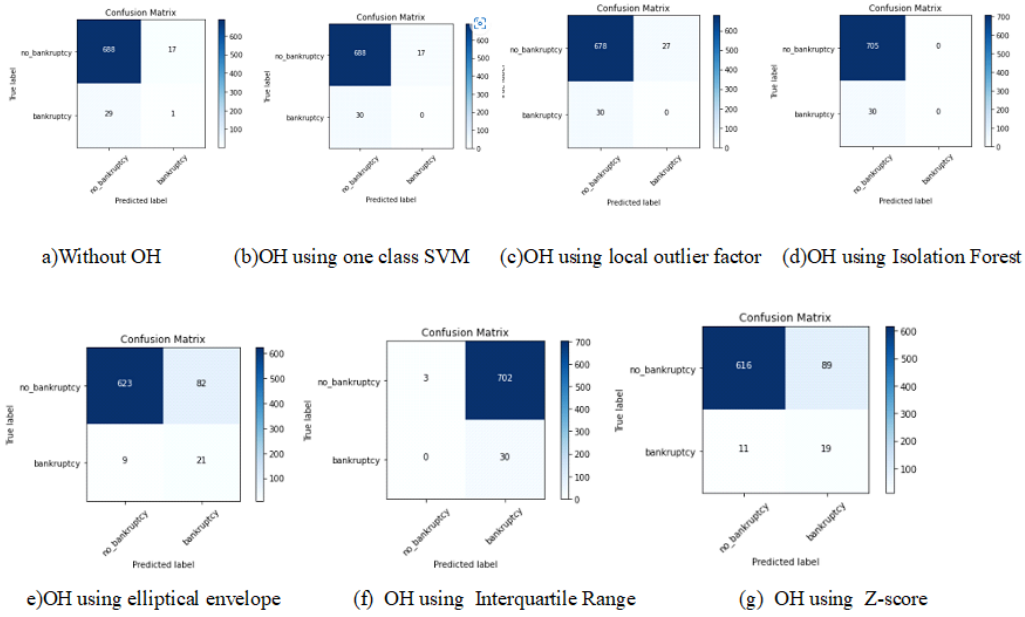
The final trained models are then produced, and their effectiveness is assessed by comparing them to the test datasets. To assess the effectiveness of the models, we employ classification performance indicators including confusion matrix, accuracy_score, f1_score, precision_score, and recall_score. We are using the Python 3.7 (TensorFlow) platform for our implementation.

5.1 Result for Dataset 1

The confusion matrices and ROC curves obtained after applying different outlier handling techniques for dataset 1, using the data generating method GAN as the balancing technique, from the ANN classifier are shown in **(a)**-**(g)** of Figures 2 and 3, from the LSTM classifier are shown in **(a)**-**(g)** of Figure 4, from the CNN classifier are shown in **(a)**-**(g)** of Figure 5 and from the ANFIS classifier are shown in **(a)**-**(g)** of Figure 6.

5.2 Result for Dataset 2 using GAN

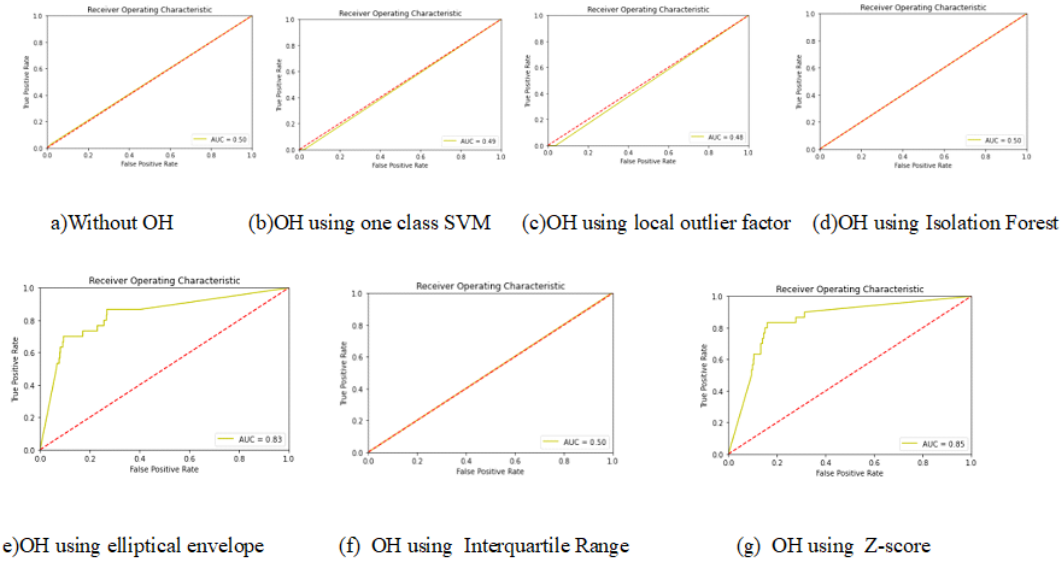
The confusion matrices and ROC curves obtained after applying different outlier handling techniques for dataset 2, using the data generating method GAN as the balancing technique, from the ANN classifier are shown in **(a)**-**(g)** of Figure 8, from the LSTM classifier are shown in **(a)**-**(g)** of Figure 9, from the CNN classifier are shown in **(a)**-**(g)** of Figure 10, and from the ANFIS classifier are shown in **(a)**-**(g)** of Figure 11.



a)Without OH (b)OH using one class SVM (c)OH using local outlier factor (d)OH using Isolation Forest

e)OH using elliptical envelope (f) OH using Interquartile Range (g) OH using Z-score

Fig 2. Confusion Matrix obtained from ANN for Dataset 1



a)Without OH (b)OH using one class SVM (c)OH using local outlier factor (d)OH using Isolation Forest

e)OH using elliptical envelope (f) OH using Interquartile Range (g) OH using Z-score

Fig 3. ROC Curve obtained from ANN for Dataset 1

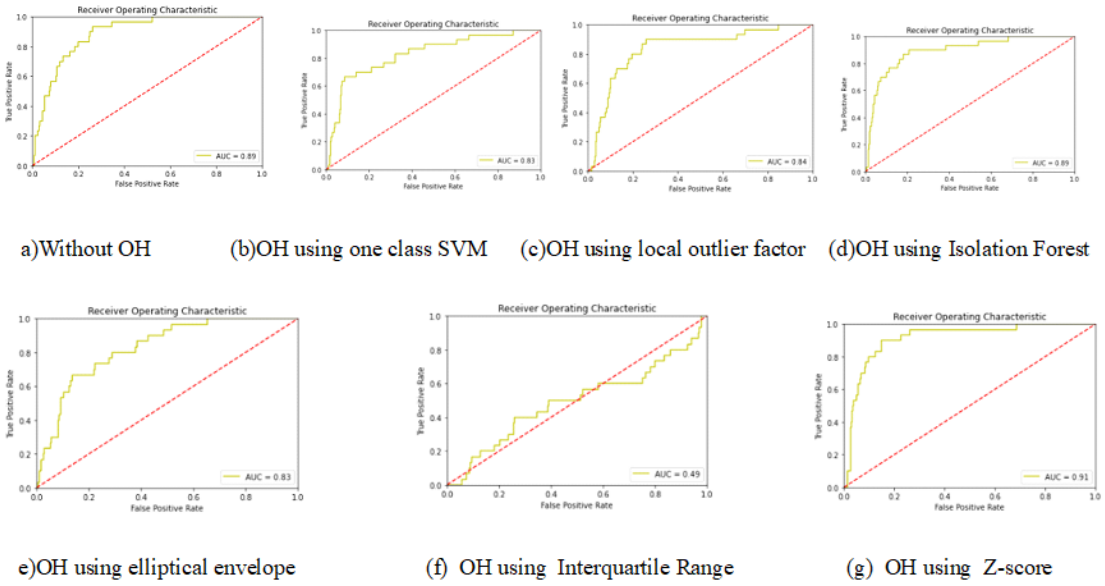


Fig 4. ROC Curve obtained from LSTM for Dataset 1

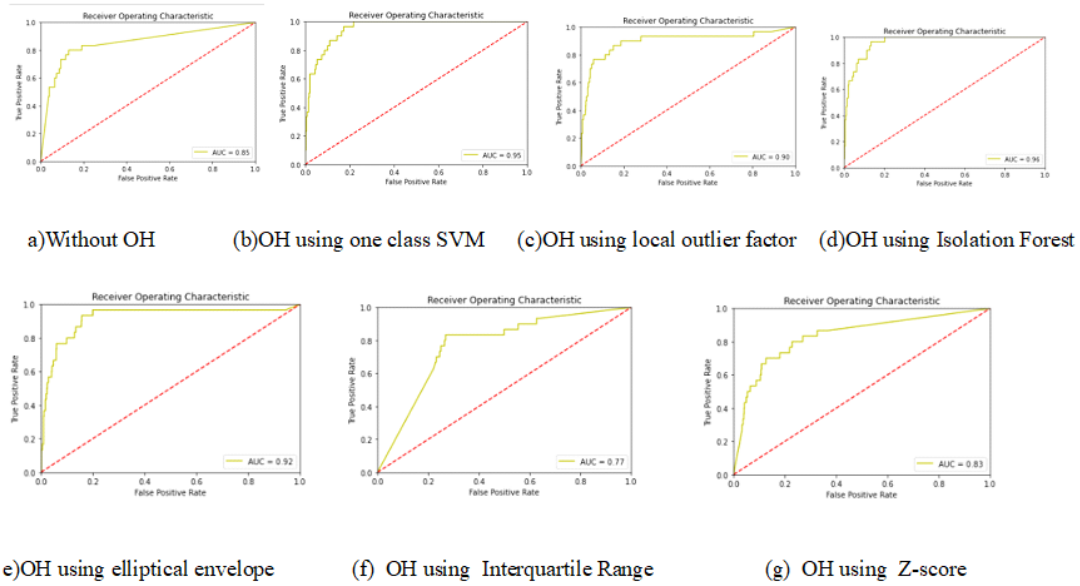
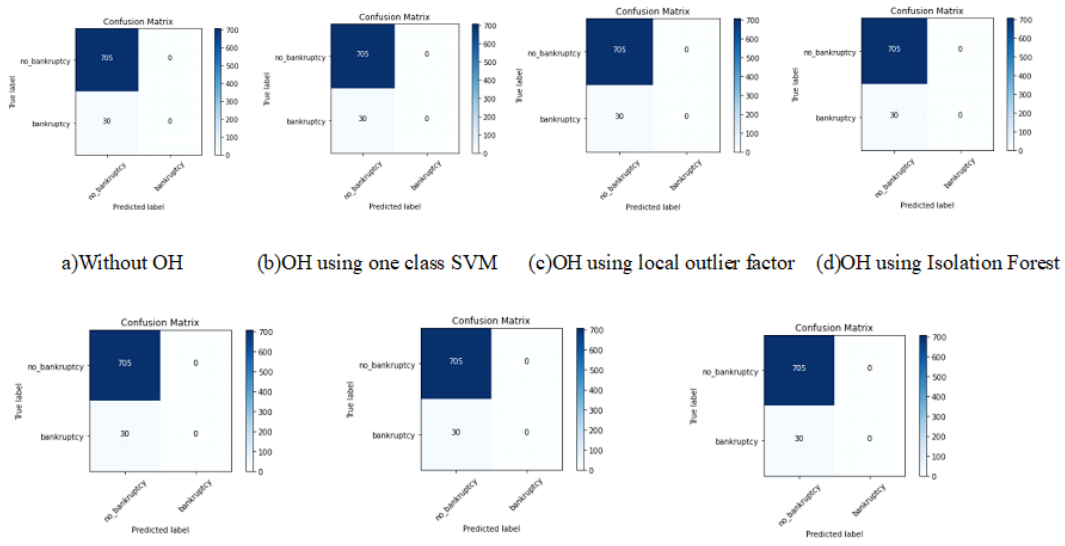


Fig 5. ROC Curve o btained from CNN for Dataset 1



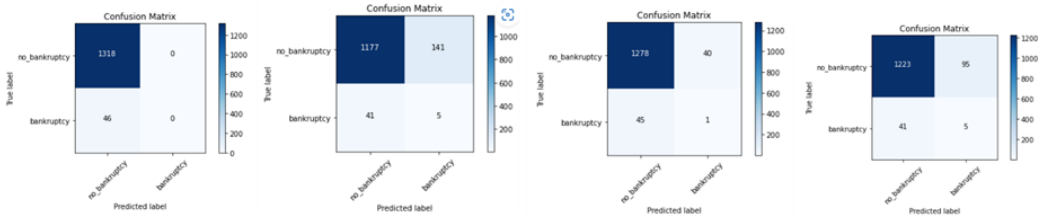
a)Without OH (b)OH using one class SVM (c)OH using local outlier factor (d)OH using Isolation Forest

(e)OH using elliptical envelope (f) OH using Interquartile Range (g) OH using Z-score

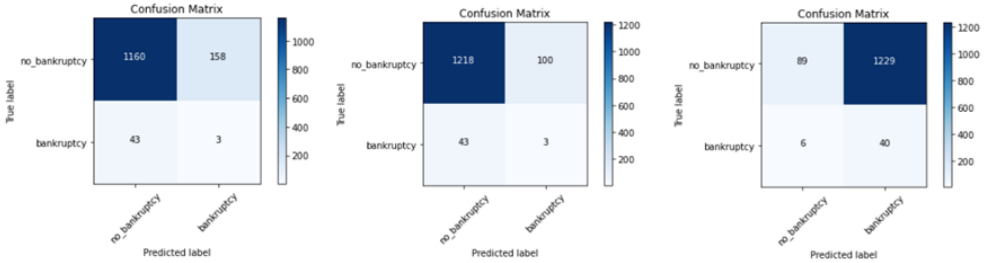
Fig 6. Confusion Matrix obtained from ANFIS for Dataset 1

Predictive Models	Performance measures	Outlier Handling Techniques						
		Without OH	OH using one class SVM	OH using local outlier factor	OH using Isolation Forest	OH using elliptical envelope	OH using IQR	OH using Z-score
ANN	Accu-score f1_score pre_score recall_score	0.8629 0.6001 0.5771 0.7515	0.9360 0.4834 0.4791 0.5	0.9224 0.4798 0.4788 0.4808	0.9591 0.4895 0.4795 0.5	0.8761 0.6238 0.5948 0.7918	0.8639 0.6001 0.5791 0.7535	0.9374 0.5046 0.5075 0.5046
LSTM	Accu-score f1_score pre_score recall_score	0.8874 0.6202 0.5909 0.7402	0.9360 0.6102 0.6055 0.6156	0.8952 0.6248 0.5949 0.7379	0.9333 0.6798 0.6466 0.7418	0.9367 0.6378 0.6285 0.6489	0.8884 0.6222 0.5929 0.7502	0.9551 0.5860 0.6638 0.5617
CNN	Accu-score f1_score pre_score recall_score	0.6251 0.6640 0.6300 0.6375	0.9428 0.7287 0.6849 0.8106	0.9115 0.6832 0.6366 0.8421	0.9333 0.7187 0.6686 0.8375	0.9442 0.7257 0.6858 0.7953	0.7156 0.5102 0.5496 0.7719	0.9265 0.6512 0.6229 0.7063
ANFIS	Accu-score f1_score pre_score recall_score	0.9701 0.4895 0.4795 0.5	0.9701 0.4895 0.4795 0.5	0,9701 0.4095 0.4795 0.5	0.9701 0.4895 0.4795 0.5	0.9701 0.4895 0.4795 0.5	0.9701 0.4895 0.4795 0.5	0.9701 0.4895 0.4795 0.5

Fig 7. Performances of different models by applying different outlier handling methods for Dataset 1

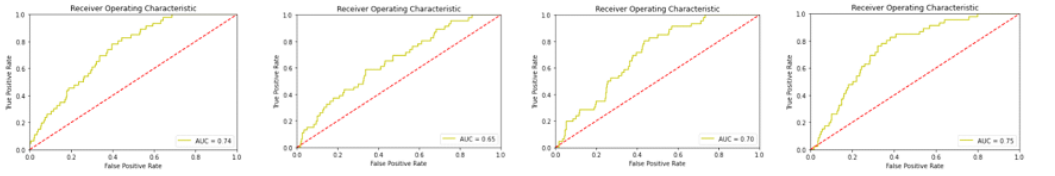


a)Without OH (b)OH using one class SVM (c)OH using local outlier factor (d)OH using Isolation Forest

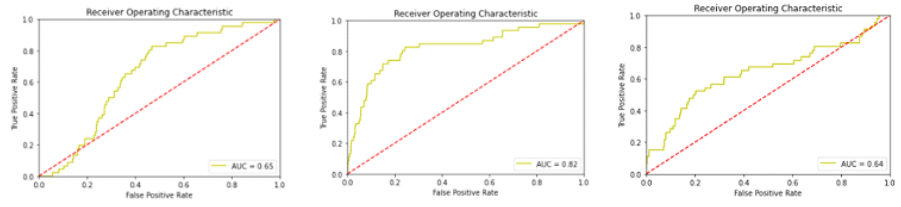


(e)OH using elliptical envelope (f) OH using Interquartile Range (g) OH using Z-score

Fig 8. Confusion Matrix obtained from ANN for Dataset 2

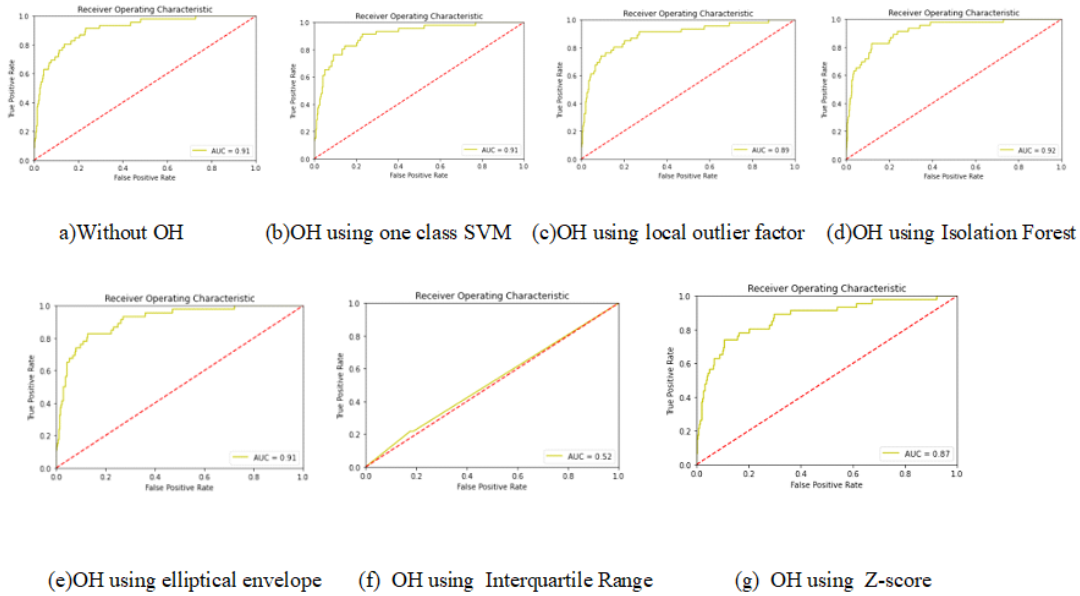


a)Without OH (b)OH using one class SVM (c)OH using local outlier factor (d)OH using Isolation Forest



(e)OH using elliptical envelope (f) OH using Interquartile Range (g) OH using Z-score

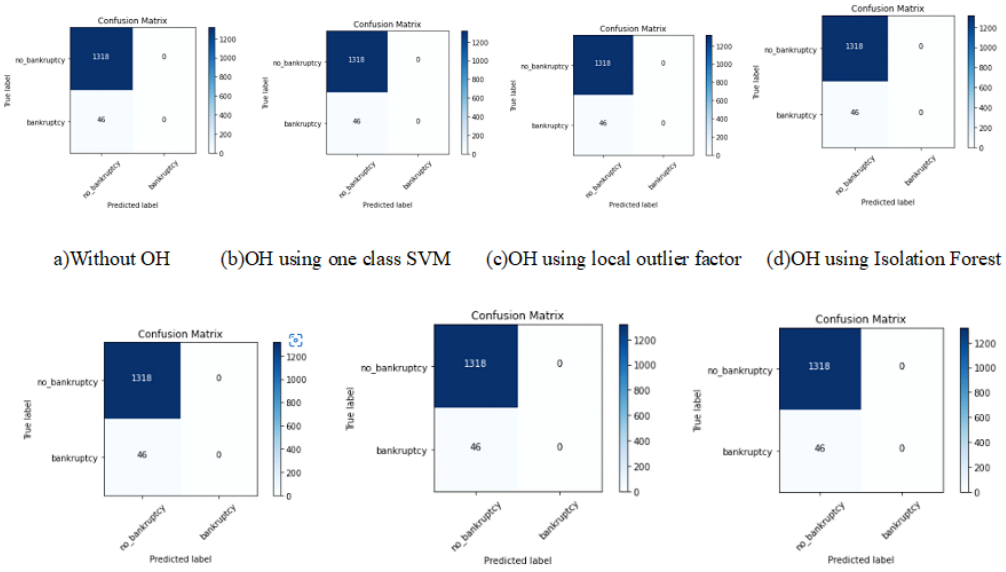
Fig 9. ROC Curve obtained from LSTM for Dataset 2



a)Without OH (b)OH using one class SVM (c)OH using local outlier factor (d)OH using Isolation Forest

(e)OH using elliptical envelope (f) OH using Interquartile Range (g) OH using Z-score

Fig 10. ROC Curve obtained from CNN for Dataset 2



a)Without OH (b)OH using one class SVM (c)OH using local outlier factor (d)OH using Isolation Forest

(e)OH using elliptical envelope (f) OH using Interquartile Range (g) OH using Z-score

Fig 11. Confusion Matrix ob tained from ANFIS for Dataset 2

Predictive Models	Performance measures	Outlier Handling Techniques						
		Without OH	OH using one class SVM	OH using local outlier factor	OH using Isolation Forest	OH using elliptical envelope	OH using IQR	OH using Z-score
ANN	Accu-score	0.8506	0.8665	0.9376	0.9002	0.9662	0.8952	0.8526
	f1_score	0.4346	0.4901	0.4954	0.5079	0.4914	0.4924	0.4746
	pre_score	0.4514	0.5002	0.4951	0.5087	0.4831	0.4975	0.4914
	recall_score	0.4526	0.5008	0.4956	0.5183	0.5	0.4946	0.4726
LSTM	Accu-score	0.3404	0.5447	0.6810	0.5351	0.6554	0.3504	0.4504
	f1_score	0.2613	0.6626	0.4545	0.3975	0.4528	0.2913	0.3913
	pre_score	0.5081	0.5126	0.5171	0.5242	0.5247	0.5181	0.5381
	recall_score	0.6118	0.5965	0.6146	0.6860	0.6748	0.6218	0.6618
CNN	Accu-score	0.7917	0.9237	0.8892	0.9076	0.9039	0.9024	0.8658
	f1_score	0.4743	0.7287	0.6251	0.6435	0.6449	0.6361	0.5985
	pre_score	0.5031	0.6206	0.5932	0.6053	0.6061	0.6001	0.5777
	recall_score	0.5146	0.7927	0.8168	0.8053	0.8244	0.8026	0.8046
ANFIS	Accu-score	0.9662	0.9662	0,9662	0.9662	0.9662	0.9662	0.9662
	f1_score	0.4914	0.4914	0.4914	0.4914	0.4914	0.4914	0.4914
	pre_score	0.4831	0.4831	0.4831	0.4831	0.4831	0.4831	0.4831
	recall_score	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Fig 12. Performances of different models by applying different outlier handling methods for Dataset 2

Figures 7 and 12 show that, for each classification model, improved prediction results were obtained after the outliers were removed. Based on these findings, we may conclude that outlier treatment improves prediction results. However, no outlier treatment technique can be declared the best since different datasets and predicting models yield varying prediction outcomes.

The highest prediction results are obtained by the GAN-ANFIS model when compared to the GAN-based ANN, LSTM, CNN, and ANFIS models. When compared to dataset2, the prediction result for dataset 1 is marginally better. Our GAN-ANFIS model's accuracy in predicting bankruptcy for dataset 1 is 0.9701, while that of GAN-ANN, GAN-LSTM, and GAN-CNN is 0.8629, 0.8874, and 0.6251. The GAN-ANFIS model's accuracy in predicting bankruptcy for dataset 2 is 0.9662, while the results for GAN-ANN, GAN-LSTM, and GAN-CNN are 0.8506, 0.3404, and 0.7917, respectively. The aforementioned tables 1 and 2 include the additional performance metrics. The accuracy of our GAN-ANFIS model is still higher than that of the GAN-ANN, GAN-LSTM, and GAN-CNN models for both datasets and every outlier treatment technique. So here we may conclude that GAN-ANFIS model is better than the hybrid models like GAN-LSTM, GAN-CNN, and GAN-ANN.

In the literature review we found some hybrid models of ANFIS like ANFIS-HB, ANFIS-GA, ANFIS-BP, and ANFIS-SA which are also used for classification purpose[8]. Among these ANFIS hybrid models ANFIS-SA is the best one as it gives the highest average accuracy rate of 96.28% whereas ANFIS-HB, ANFIS-GA and ANFIS-BP gives average accuracy rate of 76.63%, 76.63%, and 85.83%. But our GAN-ANFIS gives the highest accuracy rate of 97.01% for dataset 1 and for dataset 2 96.62%, which is also higher than the ANFIS-SA. So here we may conclude that GAN-ANFIS is the superior model for classification.

In[2] the authors used various models, but we only compare between MLSTM-3F and RZTMLSTM-3F, where in RZTMLSTM-3F, Z score is used as an outlier elimination method. When we consider about the prediction capability, here it is found out that MLSTM3F model has RMSE value 0.522 and R2 score 0.727 while RZTMLSTM-3F model has RMSE value 0.212 and R2 score 0.954. The RZTMLSTM-3F model provides the better predictive results compared to MLSTM-3F, that means after outlier elimination prediction capability increases. In[3] the interquartile range (IQR) is used to detect the outliers and kNN is used as classifier. For one dataset kNN gives accuracy score as 68.30 that time the Hybrid pre-processing + kNN gives accuracy 84.97. For other two datasets also the accuracy score of Hybrid pre-processing + kNN is more than that of kNN. In this research we also observe that our GAN based ANN, LSTM, and CNN models also increase the prediction accuracy after outlier handling. But our GAN-ANFIS model does not get affected in any way by the outlier handling techniques, that means this GAN-ANFIS model does not get affected by the outlier data, so no need of outlier handling.

6 Conclusions

Most bankruptcy databases contain unnecessary information since they are unbalanced and outlier-filled in nature. Creating trustworthy models is crucial to lowering the risk of bankruptcy. In this study, we have suggested four hybrid models for predicting bankruptcy: GAN-ANN, GAN-LSTM, GAN-CNN, and GAN-ANFIS. By producing synthesised data using the GAN (Generative Adversarial Network) approach, data set balance has been accomplished. Six outlier handling techniques have been used to compare their effects on the ability to predict bankruptcy, including one class SVM, local outlier factor, isolation forest, elliptical envelope, interquartile range IQR, and z-score. These techniques were used to select the most pertinent outlier detection techniques by utilising four different bankruptcy prediction models. GAN-ANFIS model, which shows better ability to predict bankruptcy than other three models. Surprisingly, this model exhibits resistance to anomalous data. Regardless of whether outlier handling strategies are used or not, its predicted accuracy is constant. When using each classification model, outlier control strategies produce better prediction results than when they are not used. When compared to other classification models, the GAN-ANFIS model performs better. The proposed GAN-ANFIS model's accuracy in predicting bankruptcy for dataset1 is 0.9701, whereas it is 0.9662 for dataset2. Due to the effect of the features' fuzzyness, the GAN-ANFIS model is not at all affected by the existence of outliers in the data set.

7 Author Contributions

Sasmita Manjari Nayak conducted the research, implementation, analyzed the data, and wrote the paper. Dr. Minakhi Rout guided the workflow and helps with analysis and discussion of the result.

References

1) Xu Z, Kakde D, Chaudhuri A. Automatic Hyperparameter Tuning Method for Local Outlier Factor, with Applications to Anomaly Detection. In: 2019 IEEE International Conference on Big Data (Big Data). IEEE. 2020;p. 4201–4207. Available from: https://doi.org/10.1109/BigData47090.2019.9006151.
2) Mishra S, Chawla M. A Comparative Study of Local Outlier Factor Algorithms for Outliers Detection in Data Streams. In: Emerging Technologies in Data Mining and Information Security;vol. 813 of Advances in Intelligent Systems and Computing. Springer, Singapore. 2019;p. 347–356. Available from: https://doi.org/10.1007/978-981-13-1498-8_31.
3) Urolagin S, Sharma N, Datta TK. A combined architecture of multivariate LSTM with Mahalanobis and Z-Score transformations for oil price forecasting. *Energy*. 2021;231:120963. Available from: https://doi.org/10.1016/j.energy.2021.120963.
4) Yang X, Zhou W, Shu N, Zhang H. A Fast and Efficient Local Outlier Detection in Data Streams. In: Proceedings of the 2019 International Conference on Image, Video and Signal Processing. ACM. 2019;p. 111–116. Available from: https://doi.org/10.1145/3317640.3317653.
5) Nair P, Kashyap I. Hybrid Pre-processing Technique for Handling Imbalanced Data and Detecting Outliers for KNN Classifier. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 14-16 February 2019, Faridabad, India. IEEE. 2019;p. 460–464. Available from: https://doi.org/10.1109/COMITCon.2019.8862250.
6) Sainis N, Srivastava D, Singh R. Feature classification and outlier detection to increased accuracy in intrusion detection system. *International Journal of Applied Engineering Research*. 2018;13(10):7249–7255. Available from: https://www.ripublication.com/ijaer18/ijaerv13n10_02.pdf.
7) Chen H, Ma H, Chu X, Xue D. Anomaly detection and critical attributes identification for products with multiple operating conditions based on isolation forest. *Advanced Engineering Informatics*. 2020;46:101139. Available from: https://doi.org/10.1016/j.aei.2020.101139.
8) Antonini M, Vecchio M, Antonelli F, Ducange P, Perera C. Smart Audio Sensors in the Internet of Things Edge for Anomaly Detection. *IEEE Access*. 2018;6:67594–67610. Available from: https://doi.org/10.1109/ACCESS.2018.2877523.
9) Antonini M, Vecchio M, Antonelli F, Ducange P, Perera C. Smart Audio Sensors in the Internet of Things Edge for Anomaly Detection. *IEEE Access*. 2018;6:67594–67610. Available from: https://doi.org/10.1109/ACCESS.2018.2877523.
10) Regaya Y, Fadli F, Amira A. Point-Denoise: Unsupervised outlier detection for 3D point clouds enhancement. *Multimedia Tools and Applications*. 2021;80(18):28161–28177. Available from: https://doi.org/10.1007/s11042-021-10924-x.
11) Yang J, Deng T, Sui R. An Adaptive Weighted One-Class SVM for Robust Outlier Detection. In: Proceedings of the 2015 Chinese Intelligent Systems Conference. Lecture Notes in Electrical Engineering;Springer, Berlin, Heidelberg. 2016;p. 475–484. Available from: https://doi.org/10.1007/978-3-662-48386-2_49.
12) Thomas R, Judith JE. Voting-Based Ensemble of Unsupervised Outlier Detectors. In: Advances in Communication Systems and Networks ;vol. 656 of Lecture Notes in Electrical Engineering. Springer, Singapore. 2020;p. 501–511. Available from: https://doi.org/10.1007/978-981-15-3992-3_42.
13) Feng M, Shaonan T, Chihoon L, Ling M. Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*. 2019;274(2):743–758. Available from: https://doi.org/10.1016/j.ejor.2018.10.024.
14) Kim H, Cho H, Ryu D. Corporate Bankruptcy Prediction Using Machine Learning Methodologies with a Focus on Sequential Data. *Computational Economics*. 2022;59(3):1231–1249. Available from: https://doi.org/10.1007/s10614-021-10126-5.
15) Haznedar B, Arslan MT, Kalinli A. Optimizing ANFIS using simulated annealing algorithm for classification of microarray gene expression cancer data. *Medical & Biological Engineering & Computing*. 2021;59(3):497–509. Available from: https://doi.org/10.1007/s11517-021-02331-z.
16) Ahmed IE, Mehdi R, Mohamed EA. The role of artificial intelligence in developing a banking risk index: an application of Adaptive Neural Network-Based Fuzzy Inference System (ANFIS). *Artificial Intelligence Review*. 2023;56(11):13873–13895. Available from: https://doi.org/10.1007/s10462-023-10473-9.
17) Seputra YEA. Forecasting Corporate Bankruptcy Based on Managerial Overconfidence Using the Adaptive Neuro-Fuzzy Inference System. In: Proceedings of the 3rd International Conference on Vocational Higher Education (ICVHE 2018). Advances in Social Science, Education and Humanities Research;Atlantis Press. 2020;p. 245–255. Available from: https://doi.org/10.2991/assehr.k.200331.149.

18) Talpur N, Salleh MNM, Hussain K. An investigation of membership functions on performance of ANFIS for solving classification problems. In: International Research and Innovation Summit (IRIS2017), 6–7 May 2017, Melaka, Malaysia;vol. 226 of IOP Conference Series: Materials Science and Engineering. IOP Publishing. 2017;p. 1–7. Available from: https://iopscience.iop.org/article/10.1088/1757-899X/226/1/012103/pdf.
19) Hosaka T. Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Systems with Applications*. 2019;117:287–299. Available from: https://doi.org/10.1016/j.eswa.2018.09.039.
20) Hong DS, Baik C. Generating and Validating Synthetic Training Data for Predicting Bankruptcy of Individual Businesses. *Journal of information and communication convergence engineering*. 2021;19(4):228–233. Available from: https://doi.org/10.6109/jicce.2021.19.4.228.
21) Stijven S, Minnebo W, Vladislavleva K. Separating the wheat from the chaff: on feature selection and feature importance in regression random forests and symbolic regression. In: Proceedings of the 13th annual conference companion on Genetic and evolutionary computation. 2011;p. 623–630. Available from: https://doi.org/10.1145/2001858.2002059.