IJCDS 1570979106

International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

Bankruptcy Prediction Using a GAN-based Data Augmentation Hybrid Model

Sasmita Manjari Nayak, Minakhi Rout*

School of Computer Engineering, KIIT (Deemed to be) University, Bhubaneswar, Odisha, India
*sasmita.lina@yahoo.com , minakhi.routfcs@kiit.ac.in**

Abstract: In order to lower the danger of a company failing, research in the area of bankruptcy prediction is still being done. New effective models are being developed employing a variety of cutting-edge methodologies. However, the majorities of bankruptcy databases are unbalanced and may include unnecessary data. So, creating a powerful, trustworthy model to improve prediction is always a difficult undertaking. We made the forecast in this paper in three stages. In the first stage, we concentrated on balancing the datasets using two well-known methods, SMOTETomek and GAN (Generative Adversarial Network), which generate synthesized data. Then, in the second phase, a selection of pertinent features was extracted using three wrapper-based feature selection methods: step forward feature selection, backward elimination, and recursive feature elimination, as well as five filter methods: dropping constant features, feature selection based on correlation, information gain, Chi-square test, feature importance. These three ANN, CNN, and LSTM models have been used for the third step of actual prediction. After obtaining pertinent information by feature selection from both sampling approaches, the results show that the ANN model has a better capacity for prediction than the other two predictive models. It has been demonstrated that the GAN technique outperforms the SMOTETomek with respect to all three predictive models.

Keywords: Synthetic Minority Over-Sampling Technique (SMOTE) SMOTETOmek, Feature Selection (FS), Generative Adversarial Network (GAN), Artificial neural networks (ANN), Long short term memory networks (LSTM), Convolutional neural network (CNN).

1. Introduction

When a business is unable to pay its debts, it is considered to be in financial difficulty. Bankruptcy is the final level of financial hardship. When a company stops operating altogether, it becomes bankrupt [1]. In order to prevent bankruptcy and take corrective action at the right time, corporations need effective prediction models. Predicting bankruptcy is specifically a binary classification problem. The majority of bankruptcy datasets are imbalanced in nature, and while most bankruptcy datasets have many features, only a tiny subset of these features is important for predicting bankruptcy [2]. For these reasons, pre-processing the datasets is necessary before creating models to forecast bankruptcy.

This study makes use of three datasets. These datasets have a large number of features. We now proceed with feature selection. In order to determine the optimum feature selection approach, six filter methods and three wrapper methods are employed and compared. We employ one hybrid balancing technique (SMOTETomek) and one data generating technique (GAN) to balance the datasets because they are likewise highly imbalanced in nature.

2. Literature Review

Bankruptcy is the final phase of financial distress. For bankruptcy prediction, financial institutions need effective prediction models [1]. Generally, the bankruptcy datasets are having many features and only a small subset of these features is significant for bankruptcy prediction. That's why dimensionality reduction is considered as one of the important pre-processing steps [2]. Feature selection methods were developed to choose the most appropriate features for bankruptcy prediction [3]. Several FS approaches exist they are broadly divided into three groups: filter methods, wrapper methods and embedded methods [4] and [16].

E-mail:author's email

The filter approach ranks the features according to a specific metrics, for which the performance of the classifier on the feature subset is ignored by this method [14]. The filter-based approach, has some drawbacks like: It usually chooses extensive subsets, it frequently over fits new data, and its subsequent classifier shows poor classification accuracy. For which, the classifier-based wrapper feature selection method is preferable. Wrapper methods are having better recognition rates than filter methods and can avoid over fitting problems [5]. In [5-9], the authors implement different filter methods as well as different wrapper methods for FS and found out that the wrapper methods give a better result. In [11] and [13] the authors introduced some new wrapper methods for FS and found that they are better than the other FS methods. In [10], the author uses the multi-objective evolutionary algorithm ENORA, NSGA-II, and RFE (Recursive Feature Elimination) as an FS method and found that RFE is outperformed by ENORA. In [12], five popular FS techniques, t-test, correlation matrix, stepwise regression, principal component analysis (PCA), and factor analysis (FA) are used, where t-test FS performs better than the others. In [15], the author found that the performance of the genetic algorithm as a wrapper-based FS approach is superior. In [17], three techniques were applied i.e. PCA, Select Percentile, and Sequential Feature Selection for feature selection, and found this Sequential feature selection method is giving the better accuracy value. In [18-21] different filter feature selection methods like FS with Correlation, FS on Information gain, FS on Chi-Square Test, and FS on Feature Importance are discussed. In [22] different wrapper methods are also discussed.

The unequal distributions of data among different classes in a dataset are known as the imbalanced nature of the dataset. Overall all these balancing techniques are of two types: under sampling and oversampling. In [23], the under-sampling method has been used to overcome the dataset imbalance problem. In [24], to address the imbalance problem, Synthetic Minority Over-Sampling Technique (SMOTE) is employed. In [25] a comparison takes place among different sampling methods: under sampling, oversampling, and hybrid method SMOTKTomek, and found that SMOTETomek is the bast one for balancing. The hybrid method named SMOTETomek is applied by some authors to balance the dataset and found good accuracy [26], [27]. In [28] a hybrid model named GA- ANN is introduced by combining the Genetic Algorithms (GA) and ANN.GA-ANN. In [29], to solve the data imbalance, synthetic data techniques like: CTGAN, and GAN are deployed. In [30], a comparison takes place among SMOTE, Deep SMOTE, DA-SMOTE, GAN, and Borderline SMOTE.

In this literature review, we got different types of classification methods. In [31], it is found that deep learning algorithms outperform traditional models in predicting bankruptcy based on textual data. The prediction capability of RNN and LSTM methodologies can outperform all traditional machine learning[37] models [32,33]. Researchers compared logistic regression and ANNs models and found out that the ANN gives better results than the logistic regression when a larger dataset is used for testing [34], [35].

Through this literature review we found that, in the case of FS, most of the research papers only worked on some specific FS techniques. In this research, an extensive investigation is carried out to investigate the impact of feature selection by applying four different filter methods and three different wrapper-based FS approaches. and try to find out which one is more suitable one for bankruptcy datasets. Through this literature review, we also got different balancing techniques. We observed that among various sampling methods the hybrid method SMOTETomek is yielding better predictive result and simultaneously it is also discovered that deep learning based data generation technique Generative Adversial Network (GAN) is used widely by the researcher nowadays. Hence, we proposed to compare between the statistical method SMOTETomek and deep learning method GAN to generate synthesis data in order to balance the dataset considered here for this study. In this literature review we also found that deep learning methods outperforms than the traditional machine learning models. After the balancing the dataset, the relevant features will be extracted by applying different feature selection methods and compare the predictive results, by taking machine learning models: CNN, LSTM, and ANN. Here our aim is to find out the composition of balancing technique, feature selection technique, and classification model which outperforms in the domain of bankruptcy prediction.

3. METHODOLOGY

Figure 1 provides a step-by-step visual representation of the bankruptcy prediction process, while Figure 2 displays the detailed architecture of the GAN. There are two categories for the datasets: 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. Two types of balancing techniques are applied to the imbalanced training datasets: one is the hybrid method called SMOTETomek, which applies Tomek linkages as a data cleaning technique to the SMOTE over-sampled training set; the other method is called GAN (Generative Adversarial Network), which generates the necessary number of minority class data to balance the dataset.

Next, we go on to the second round of preprocessing operations, where we pick and choose whatever features are most important or required and remove the remainder. In this case, we employ both wrapper and filter approaches. In this study, five distinct filer techniques are examined: feature selection with correlation, FS on information gain, FS on Chi-square test, and FS based on feature importance. First, we use the DropConstantFeatures() method to remove constant features in order to start the feature selection process. Next, we employ more feature selection techniques one after the other. We go on to wrapper feature selection techniques after filter methods. Three different kinds of wrapper approaches are used here: backward elimination, step forward feature selection, and random forest classifier in a recursive manner to exclude features. Following the conclusion of the feature selection process, the classification models are implemented. Here, we're using ANN, CNN, and LSTM—three different kinds of categorization models—to forecast bankruptcy.

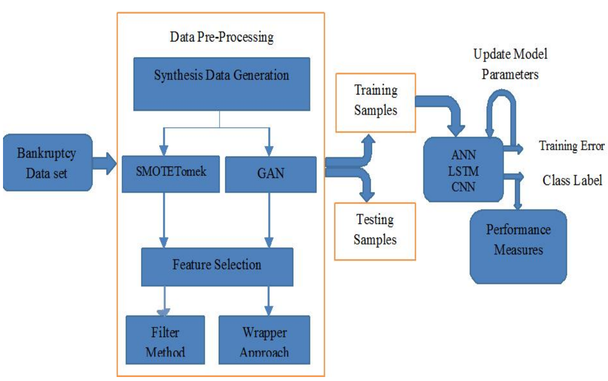


Fig. 1 The workflow of proposed bankruptcy prediction process

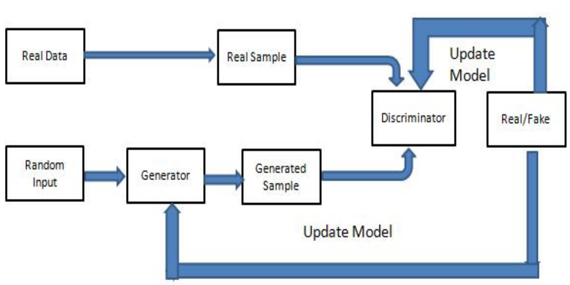


Fig. 2 The architecture of GAN.

The specification of each classifier is described as follows: ANN model is having one input layer, two hidden layers, and one output layer. Here the number of neurons in the input layer is changed from time to time, as the number of features in the dataset which is used for prediction is changed, but the output layer has 2 neurons always, as we need two outputs i.e. bankrupt and non-bankrupt. Here one of the hidden layers is having 20 neurons whereas the other is having 10 neurons.

The LSTM [37] model is sequentially dense. Here the first LSTM layer is having 80 neurons. We are using the input shape as the number of features in the dataset which is used for prediction. One more layer was then added. Finally, the output layer having 2 neurons for two output classes i.e. bankrupt and non-bankrupt. We start with the LSTM layer, then a few Dropout layers are added, to avoid over fitting. We specify 0.2 for the Dropout layers, which means 20% of the layers, will be removed. The Dense layer is next added, which specifies a single output unit. Here sigmoid activation function is used.

The CNN model has Conv1D (). So it is having a one-dimensional convolution layer. In the first Conv1D () layer, over the dataset, we apply 128 filters, with the convolutional window having a size of 3. The input_shape parameter takes the value as the number of features in the dataset to be predicted. Finally, four layers of hidden neurons with rectified linear activation function (ReLU) activation function are used. Here pool_size is taken as 2. After that, the output comes, which is a dense layer with 2 neurons as always only two predicted value that is either 0 or 1, we need. Here the softmax activation function is used as, it ranges from 0 to1, which makes it easy to predict a binary value as an output. In this model Flatten () is used to convert the data into a one-dimensional array for further processing to the next layer.

4. Implementation and Result Discussion

We use the same functions and settings for each model's compilation, such as the loss function "sparse_categorical_crossentropy," and we train our models using an Adam optimizer with a learning rate of 0.058 to optimise the loss function. The model is to be run with a batch size of 10 and an epoch count of 80. We are using the Python 3.7 (TensorFlow) platform for our implementation.

The preprocessed training dataset is used to train the models in classification after preprocessing. Subsequently, the trained model is finalised and tested against the test dataset to assess its performance. Confusion matrices, ROC curves, accuracy, F1 scores, precision, and recall are used to validate models and assess their efficacy.

A. Dataset Description

In this research for bankruptcy prediction, three related datasets are used, which are the Taiwan Stock Exchange, financial distress prediction, and US Bankruptcy Prediction datasets. The numbers of features and ratios of the bankrupt and non-bankrupt cases of these datasets are different to each other. The detailed information about datasets are listed in Table 1 [15].

TABLE 1: Detail Information about Datasets

Name of Dataset	No. of features	Total No. of Instances	Non-Bankrupt Instances	Bankrupt Instances
financial distress prediction	87	3682	3546	136
Company bankruptcy prediction	96	6816	6599	220
US Bankruptcy Prediction	15	92872	92872	558

B. Result for Dataset 1 using SMOTEtomek

The performance matrices are obtained for dataset 1 using SMOTETomek balancing techniques with four different filter methods and three different wrapper methods by applying to three different machine learning models ANN, LSTM, CNN and the comparison is presented in the resultant Table 2.

TABLE 2: Performances of different models by applying different FS technique for Dataset 1 using SMOTETomek

Predicti-ve Models	Performance measures	Feature Selection Techniques							
		With Out FS	Filter Methods				Wrapper Methods		
			FS With Correlation	FS on Information gain	FS on Chi square Test	FS on Feature Importance	Step Forward FS	Backward Elimination FS	Recursive Feature Elimination
ANN	Accu-score f1_score pre_score recall_score	0.7001 0.4322 0.4929 0.4567	0.9472 0.4864 0.4828 0.4901	0.9662 0.4914 0.4831 0.5	0.9662 0.4914 0.4831 0.5	0.9090 0.6458 0.6069 0.8060	0.6884 0.4273 0.4922 0.4506	0.9662 0.4914 0.4831 0.5	0.9501 0.5512 0.5627 0.5441
LSTM	Accu-score f1_score pre_score recall_score	0.4252 0.3384 0.5233 0.6711	0.4457 0.3463 0.5172 0.6292	0.6356 0.3961 0.4811 0.3708	0.5557 0.4005 0.5161 0.6232	0.0513 0.051 0.5171 0.5091	0.6385 0.4554 0.5343 0.75	0.9002 0.6037 0.5779 0.7176	0.1686 0.1600 0.5194 0.5698
CNN	Accu-score f1_score pre_score recall_score	0.9010 0.6341 0.5987 0.8019	0.5637 0.4108 0.5236 0.6798	0.6356 0.3961 0.4811 0.3708	0.7690 0.4995 0.5273 0.6497	0.909 0.6591 0.6158 0.8480	0.7346 0.5058 0.5428 0.7682	0.8577 0.5789 0.5651 0.7585	0.9039 0.6449 0.6061 0.8244

C. Result for Dataset 1 using GAN

The ROC curves obtained for the four feature selection methods which are giving better prediction result than others feature selection methods for dataset 1 with GAN as a balancing technique and ANN classifier are shown in Fig. 3, with LSTM classifier are shown in Fig. 4, with CNN classifier are shown in Fig. 5

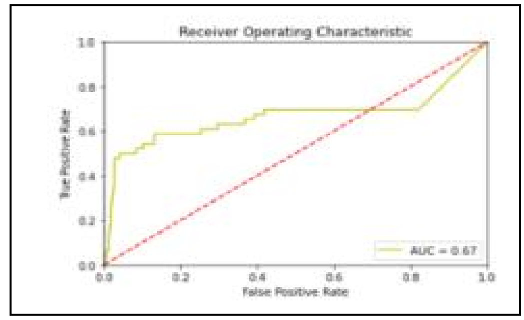


Fig. 3 ROC Curve obtained from ANN for Dataset 1 using GAN with Feature Importance.

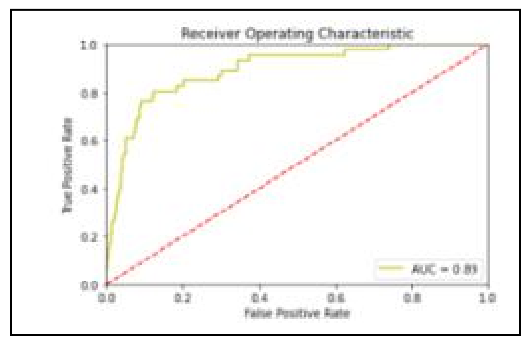


Fig. 4 ROC Curve obtained from LSTM for Dataset 1 using GAN with Feature Correlation.

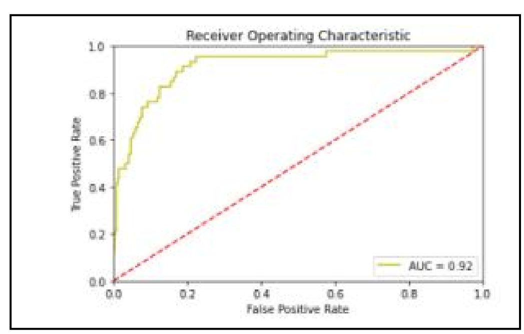


Fig. 5 ROC Curve obtained from CNN for Dataset 1 using GAN using Feature Importance.

The performance matrices are obtained for dataset 1 using GAN as a balancing techniques with four different filter methods and three different wrapper methods by applying to three different machine learning models ANN, LSTM, CNN and the comparison is presented in the resultant Table 3.

Table 3: Performances of different models by applying different FS technique for Dataset 1 with GAN.

Predicti-ve Models	Performance measures	Feature Selection Techniques							
		With Out FS	Filter Methods				Wrapper Methods		
			FS With Correlation	FS on Information gain	FS on Chi square Test	FS on Feature Importance	Step Forward FS	Backward Elimination FS	Recursive Feature Elimination
ANN	Accu-score f1_score pre_score recall_score	0.9464 0. 4984 0.5003 0.5002	0.9662 0.4914 0.4831 0.5	0.8848 0.6227 0.5920 0.8250	0.9662 0.4914 0.4831 0.5	0.9567 0.6154 0.6379 0.5999	0.9530 0.5976 0.61S1 0.4506	0.9215 0.5373 0.5315 0.5502	0.9318 0.6216 0.5975 0.6709
LSTM	Accu-score f1_score pre_score recall_score	0.5967 0.4246 0.6076 0.6654	0.9032 0.6471 0.6076 0.8345	0.1972 0.1828 0.5149 0.5636	0.8159 0.4957 0.5135 0.5586	0.1620 0.1537 0.5098 0.5349	0.3658 0.2902 0.4961 0.4725	0.6568 0.4804 0.5220 0.6546	0.2360 0.2081 0.4940 0.4683
CNN	Accu-score f1_score pre_score recall_score	0.9083 0.6446 0.6061 0.8057	0. 8973 0. 6291 0. 5954 0.8000	0.8892 0.6312 0.5974 0.8379	0.6268 0.4273 0.5120 0.5866	0.8929 0.6329 0.5983 0.8292	0.8247 0.5588 0.5578 0.7729	0.6942 0.5789 0.5352 0.7369	0.8892 0.6121 0.5842 0.7748

By observing Table 2 and Table 3, we may here conclude that without FS, CNN gives better performance and outperformed the other two models on both sampling techniques, whereas LSTM shows the worst predictive result, but after doing FS, ANN gives better performance and outperformed other two models on both sampling techniques for dataset 1. By using different FS techniques, we got different prediction values. Here we may not take one FS method as the best one for dataset 1, but after doing FS the bankruptcy prediction rate increases as compared to without FS. Here after balancing the dataset using GAN gives a better bankruptcy prediction result as compared to SMOTETomek.

D. Result for Dataset 2 using SMOTEtomek

The performance matrices are obtained for dataset 2 using SMOTETomek balancing techniques with four

Table 4: Performances of different models by applying different FS technique for Dataset 2 using SMOTETomek.

Predicti-ve Models	Performance measures	Feature Selection Techniques							
		With Out FS	Filter Methods				Wrapper Methods		
			FS With Correlation	FS on Information gain	FS on Chi square Test	FS on Feature Importance	Step Forward FS	Backward Elimination FS	Recursive Feature Elimination
ANN	Accu-score f1_score pre_score recall_score	0.9482 0.4867 0.4793 0.4943	0.9102 0.6861 0.6382 0.8574	0.8176 0.5837 0.5768 0.8251	0.9578 0.5195 0.6468 0.5152	0.8571 0.6169 0.5962 0.8297	0.9278 0.6455 0.6202 0.6911	0.0408 0.0392 0.0204 0.5	0.8666 0.6227 0.5952 0.8187
LSTM	Accu-score f1_score pre_score recall_score	0.8231 0.5999 0.5880 0.8574	0.9102 0.5318 0.5649 0.6292	0.8258 0.6098 0.5949 0.9092	0.8666 0.6404 0.6082 0.8826	0.9102 0.6211 0.5975 0.8734	0.9102 0.670 0.6277 0.8095	0.0408 0.6227 0.5952 0.8187	0.8149 0.6169 0.5727 0.8078
CNN	Accu-score f1_score pre_score recall_score	0.7360 0.5264 0.5563 0.7985	0.8517 0.6069 0.5862 0.8109	0.8204 0.5517 0.5516 0.6989	0.8068 0.5547 0.5572 0.7397	0.8068 0.5855 0.5815 0.8673	0.8247 0.5588 0.5578 0.7729	0.8952 0.6309 0.5992 0.7539	0.8448 0.6449 0.5953 0.8712

E. Result for Dataset 2 using GAN

The ROC curves obtained for the four feature selection methods which are giving better prediction result than others feature selection methods for dataset 2 with GAN as a balancing technique and ANN classifier are shown in Fig. 6, with LSTM classifier are shown in Fig. 7, with CNN classifier are shown in Fig. 8.

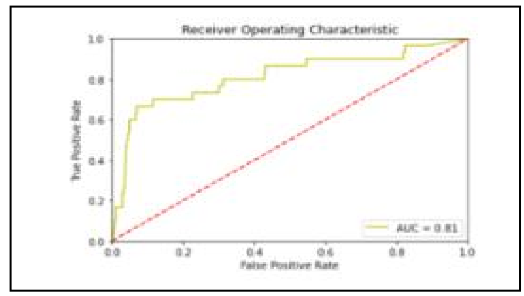


Fig. 6 ROC Curve obtained from ANN for Dataset 2 using GAN With Correlation

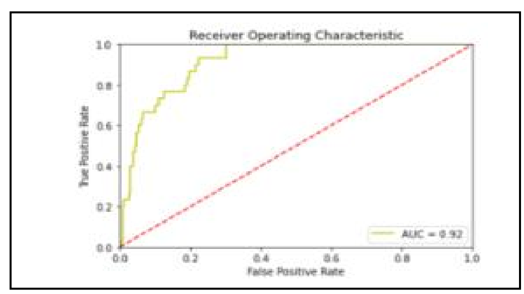


Fig. 7 ROC Curve obtained from LSTM for Dataset 2 using GAN Recursive Feature Elimination.

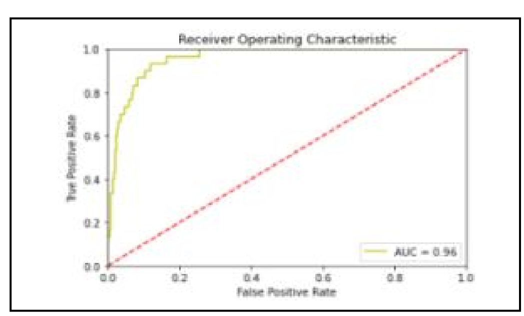


Fig. 8 ROC Curve obtained from CNN for Dataset 2 using GAN with Feature Importance.

The performance matrices are obtained for dataset 2 using SMOTETomek balancing techniques with four different filter methods and three different wrapper methods by applying to three different machine learning models ANN, LSTM, CNN and the comparison is presented in the resultant Table 5.

TABLE 5: Performances of different models by applying different FS technique for Dataset 2 with GAN

Predictive Models	Performance measures	Feature Selection Techniques							
		With Out FS	Filter Methods				Wrapper Methods		
			FS With Correlation	FS on Information gain	FS on Chi square Test	FS on Feature Importance	Step Forward FS	Backward Elimination FS	Recursive Feature Elimination
ANN	Accu-score	0.9374	0.9401	0.9414	0.9605	0.0408	0.9591	0.8326	0.9591
	f1_score	0.5046	0.6051	0.5071	0.5221	0.0392	0.4895	0.5756	0.4895
	pre_score	0.5075	0.6087	0.5131	0.9802	0.0204	0.4795	0.5668	0.4795
	recall_score	0.5046	0.6017	0.5067	0.5166	0.5	0.5	0.7531	0.5
LSTM	Accu-score	0.8952	0.9319	0.9659	0.9496	0.9414	0.9142	0.9632	0.9469
	f1_score	0.6248	0.6145	0.6149	0.6835	0.7123	0.5338	0.6765	0.7260
	pre_score	0.5949	0.6033	0.8102	0.6807	0.6749	0.5300	0.7924	0.6913
	recall_score	0.7379	0.6294	0.6790	0.8826	0.7780	0.5404	0.6297	0.7808
CNN	Accu-score	0.6251	0.9102	0.8775	0.7387	0.9251	0.9197	0.8027	0.9319
	f1_score	0.6640	0.6585	0.5517	0.7251	0.7128	0.6680	0.5297	0.7041
	pre_score	0.6300	0.6200	0.5883	0.6775	0.6598	0.6295	0.5390	0.6592
	recall_score	0.6375	0.7776	0.7606	0.8244	0.8652	0.7666	0.6578	0.8049

By observing Table 4 and Table 5, we conclude that before FS as well as after FS, ANN gives better performance and outperformed the other two models on both sampling techniques, whereas CNN shows the worst predictive result. By using different FS techniques, we got different prediction values. Here we may not take one FS method as the best one for dataset 2, but after doing FS the bankruptcy prediction rate increases as compared to without FS. Here also after balancing the dataset using GAN gives a better bankruptcy prediction result as compared to SMOTETomek.

F. Result for Dataset 3 using SMOTETomek

The performance matrices are obtained for dataset 3 using SMOTETomek balancing techniques with four different filter methods and three different wrapper methods by applying to three different machine learning models ANN, LSTM, CNN and the comparison is presented in the resultant Table 6.

Table 6: Performances of different models by applying different feature selection technique for Dataset 3 using SMOTETomek

Predictive Models	Performance measures	Feature Selection Techniques							
		With Out FS	Filter Methods				Wrapper Methods		
			FS With Correlation	FS on Information gain	FS on Chi square Test	FS on Feature Importance	Step Forward FS	Backward Elimination FS	Recursive Feature Elimination
ANN	Accu-score	0.0060	0.9102	0.8257	0.8373	0.8885	0.0060	0.9938	0.9939
	f1_score	0.0060	0.4791	0.4770	0.4827	0.5028	0.0060	0.4984	0.4984
	pre_score	0.0030	0.5130	0.5119	0.5132	0.5158	0.0030	0.4969	0.4969
	recall_score	0.5	0.8120	0.7892	0.8038	0.7636	0.5	0.4999	0.5
LSTM	Accu-score	0.8815	0.8615	0.8767	0.9099	0.8911	0.8916	0.9092	0.9939
	f1_score	0.5022	0.4948	0.4927	0.5124	0.5039	0.5020	0.5094	0.4993
	pre_score	0.5880	0.5145	0.5119	0.5178	0.5160	0.5148	0.5161	0.5130
	recall_score	0.5119	0.7760	0.7137	0.7436	0.7605	0.7387	0.7840	0.7083
CNN	Accu-score	0.8279	0.8337	0.8181	0.8579	0.8885	0.8591	0.7804	0.8288
	f1_score	0.4776	0.4793	0.4742	0.4915	0.4873	0.4907	0.4589	0.4786
	pre_score	0.5119	0.5118	0.5117	0.5147	0.5133	0.5578	0.5099	0.5123
	recall_score	0.7859	0.7756	0.7942	0.8010	0.7844	0.7840	0.7840	0.7951

G. Result for Dataset 3 using GAN

The ROC curves obtained for the four feature selection methods which are giving better prediction result than others feature selection methods for dataset 3 with GAN as a balancing technique and ANN classifier are shown in Fig. 9 and 10, with LSTM classifier are shown in Fig. 11, with CNN classifier are shown in Fig. 12 and 13.

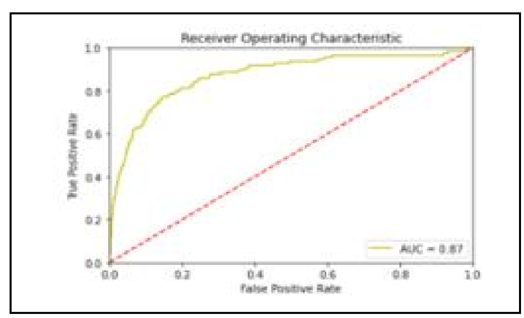


Fig. 9 ROC Curve obtained from ANN for Dataset 3 using GAN with Information gain.

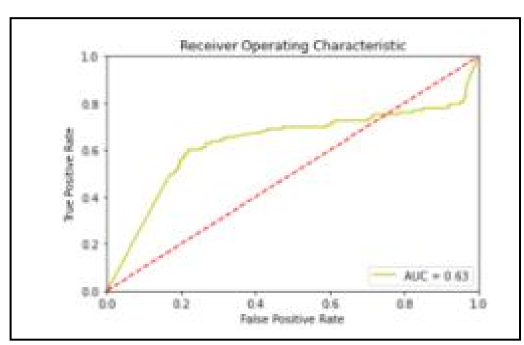


Fig. 10 ROC Curve obtained from ANN for Dataset 3 using GAN using Chi-square

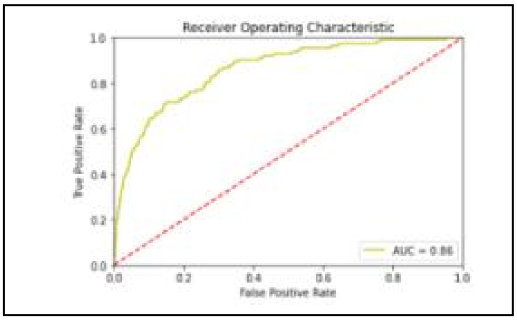


Fig. 11 ROC Curve obtained from LSTM for Dataset 3 using GAN with Feature importance.

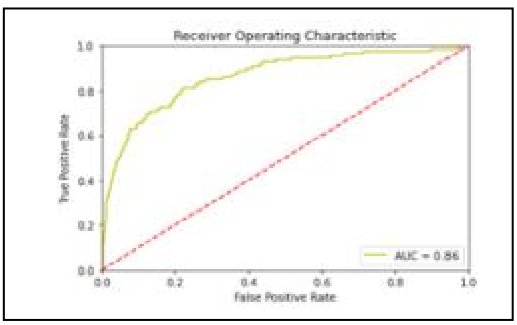


Fig. 12. ROC Curve obtained from CNN for Dataset 3 using GAN with Information gain.

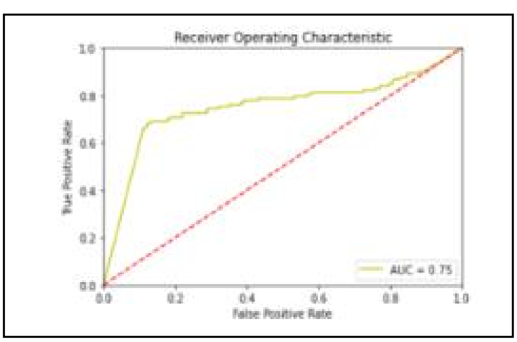


Fig. 13 ROC Curve obtained from CNN for Dataset 3 using GAN using Ch-isquare Test.

The performance matrices are obtained for dataset 3 using SMOTETomek balancing techniques with four different filter methods and three different wrapper methods by applying to three different machine learning models ANN, LSTM, CNN and the comparison is presented in the resultant Table 7.

Table 7: Performances of different models by applying different FS technique for Dataset 3 with GAN

Predicti-ve Models	Performance measures	Feature Selection Techniques							
		With Out FS	Filter Methods				Wrapper Methods		
			FS With Correlation	FS on Information gain	FS on Chi square Test	FS on Feature Importance	Step Forward FS	Backward Elimination FS	Recursive Feature Elimination
ANN	Accu-score	0.8477	0.9939	0.9414	0.9605	0.9938	0.8242	0.8529	0.0061
	f1_score	0.4881	0.4984	0.4899	0.4136	0.4984	0.4765	0.4894	0.0060
	pre_score	0.5146	0.4969	0.5147	0.5044	0.4969	0.5120	0.5143	0.5030
	recall_score	0.8178	0.5	0.8115	0.6639	0.4999	0.7928	0.8029	0.5
LSTM	Accu-score	0.9394	0.9359	0.9640	0.7992	0.9938	0.9647	0.8529	0.9590
	f1_score	0.5346	0.5347	0.5466	0.4645	0.5455	0.5561	0.5553	0.5392
	pre_score	0.5262	0.5267	0.5308	0.5094	0.5304	0.5365	0.5360	0.5266
	recall_score	0.7699	0.7699	0.6790	0.7538	0.7780	0.7008	0.7046	0.6671
CNN	Accu-score	0.7255	0.7180	0.8775	0.8081	0.7953	0.6580	0.8486	0.7720
	f1_score	0.4374	0.4344	0.4645	0.4681	0.5021	0.4103	0.4839	0.4546
	pre_score	0.5081	0.5078	0.5103	0.5099	0.5056	0.5062	0.5117	0.5089
	recall_score	0.7695	0.7657	0.7826	0.7583	0.7629	0.7356	0.7523	0.7621

By observing Table 6 and Table 7, we may here conclude that without FS, LSTM gives better performance and outperformed the other two models on both sampling techniques, whereas CNN shows the worst predictive result, but after doing FS, ANN gives better performance and outperformed other two models on both sampling techniques for dataset 3. By using different FS techniques we got different prediction values. Here we may not take one FS method as the best one for dataset 3, but after doing FS the bankruptcy prediction rate increases as compared to without FS. Here also after balancing the dataset using GAN gives a better bankruptcy prediction result as compared to SMOTETomek.

From the above resultant Tables 2-7, we can easily compare our three predictive models and two sampling methods, and feature selection methods. In each predictive model, the performance measures like: accuracy rate, f1_score, precision_score, and recall_score have remained high where the GAN (Generative Adversarial Network) is used as a balancing technique as compared to the hybrid balancing method SMOTETOmek and simultaneously the confusion matrix shows that the number of correct predictive cases of both non-bankrupt cases as well as bankrupt cases is also high. When we consider feature selection methods after FS always prediction rate increases as compared to the prediction rate before FS. After doing FS we find out that the ANN classifier gives the highest prediction value as compared to the other two classifiers i.e. LSTM and CNN, but without FS for different datasets different classifiers give the best result. But we may not take any FS method as the best for any classifier or any dataset as the classifiers are giving randomly different prediction rates for different FS methods.

5. CONCLUSIONS

In this study, we first contrast the impact of GAN on data balancing with that of the SMOTETomek hybrid sampling technique. We then gave feature selection some thought. Here, we take a look at about five filter approaches and three wrapper approaches to use three distinct bankruptcy prediction models—ANN, CNN, and LSTM—to choose the most pertinent features. The investigation yielded the following findings:

When compared to the hybrid SMOTETomek approach, the prediction result obtained from the dataset balanced by the GAN method is superior.

When feature selection is used, prediction outcomes are superior than when it is not used.

Any feature selection method cannot be declared the best because it produces various prediction outcomes for different datasets and balancing techniques.

After obtaining pertinent features through feature selection, the ANN model outperforms the other two predictive models in terms of prediction accuracy.

In this work, we take into account the dataset's imbalance while concurrently concentrating on feature selection to obtain significant or pertinent characteristics; we do not, however, pay attention to the outlier data. Therefore, in order to produce a robust prediction, we will address outliers at the model level rather than during the pre-processing stage in our future research.

REFERENCES

[1] G. Jandaghi, A. Saran., R. Rajaei, A. Ghasemi, & R. Tehrani, “Identification of the Most Critical Factors in Bankruptcy Prediction and Credit Classification of Companies”, Iranian Journal of Management Studies, Vol.14, No.4, 2021, pp. 817-834.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] A. Mustaqeem, S. M Anwar, M. Majid, & A. R. Khan, “Wrapper method for feature selection to classify cardiac arrhythmia”, In 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, pp. 3656-3659.K. Elissa, “Title of paper if known,” unpublished.

[3] N. Hussain, M. A. Khan, M. Sharif, “A deep neural network and classical features based scheme for objects recognition: an application for machine inspection”, Multimedia Tools and Applications, 2020, pp. 1-23

[4] M. Wagle, Z. Yang, & Y. Benslimane, “Bankruptcy prediction using data mining techniques”, In 2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES) IEEE, 2017, pp. 1-4.

[5] Y. D. Zhang and L. N. Wu. "Bankruptcy prediction by genetic ant colony algorithm", Advanced Materials Research. Vol. 186, 2011, pp. 459-463.

[6] C. F. Tsai, K. L. Sue, Y. H. Hu and A. Chiu, "Combining feature selection, instance selection, and ensemble classification techniques for improved financial distress prediction", Journal of Business Research, Vol. 130, 2021,pp. 200-209.

[7] D. T. Pham, M. Mahmuddin, S. Otri and H. Al-Jabbouli, "Application of the bees algorithm to the selection features for manufacturing data", 2007, pp. 1-6.

[8] E. I. Altman, M. Iwanicz-Drozdowska, E. K. Laitinen and A. Suvas, “A race for long horizon bankruptcy prediction”. Applied Economics, Vol. 52, No. 37, 2020, pp.4092-4111.

[9] D. Liang, C. F. Tsai, and H. T. Wu. "The effect of feature selection on financial distress prediction", Knowledge-Based Systems, Vol. 73, 2015, pp. 289-297.

[10] F. Jiménez, G. Sánchez, J. M. García, G. Sciavicco and l. Miralles "Multi-objective evolutionary feature selection for online sales forecasting", Neurocomputing, Vol. 234, 2017, pp. 75-92.

[11] J. B. Yang, K. Q. Shen, C. J. Ong, X. P. Li, "Feature selection for MLP neural network: the use of random permutation of probabilistic outputs", IEEE Transactions on Neural Networks, Vol. 20, No. 12, 2009, pp. 1911-1922.

[12] C. F. Tsai, “Feature selection in bankruptcy prediction", Knowledge-Based Systems, Vol. 22, No. 2, 2009, pp. 120-127.

[13] M. M. Kabir, M. M. Islam, K. Murase, "A new wrapper feature selection approach using neural network", Neurocomputing, Vol. 73, No. 16-18, 2010, pp. 3273-3283.

[14] S. Das, P. K. Singh, S. Bhowmik, R. Sarkar, "A harmony search based wrapper feature selection method for holistic bangla word recognition", Procedia Computer Science, Vol. 89 , 2016, pp. 395-403.

[15] W. C. Lin, Y. H. Lu, C. F. Tsai, "Feature selection in single and ensemble learning based bankruptcy prediction models", Expert Systems, Vol. 36, No. 1, 2019, e12335.

[16] I. Guyon, S. Gunn, M. Nikravesh, L.A. Zadeh “ Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing) Springer-Verlag (2006)

[17] A. Tabbakh, "Bankruptcy Prediction using Robust Machine Learning Model", Turkish Journal of Computer and Mathematics Education (TURCOMAT), Vol. 12, No. 10 2021, pp. 3060-3073.

[18] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection--a comparative study", Lecture notes in computer science, Vol. 4881, 2007, pp. 178-187.

[19] S. Lei, "A feature selection method based on information gain and genetic algorithm", International conference on computer science and electronics engineering, IEEE, Vol. 2, 2012, pp.355-358.

[20] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM", Journal of King Saud University-Computer and Information Sciences, Vol. 29, No. 4, 2017, pp. 462-472.

[21] S. Stijven, W. Minnebo and K. Vladislavleva, "Separating the wheat from the chaff: on feature selection and feature importance in regression random forests and symbolic regression", Proceedings of the 13th annual conference companion on Genetic and evolutionary computation, 2011, pp. 623-630.

[22] X. Deng, Y. Li, J. Weng and J. Zhang, "Feature selection for text classification: A review", Multimedia Tools and Applications, Vol. 78, 2019, pp. 3797-3816.

[23] Chen, Mu-Yen. "Predicting corporate financial distress based on integration of decision tree classification and logistic regression." Expert systems with applications 38.9 (2011): 11261-11272.

[24] T. Hosaka, "Bankruptcy prediction using imaged financial ratios and convolutional neural networks", Expert systems with applications, Vol. 117, 2019, pp. 287-299.

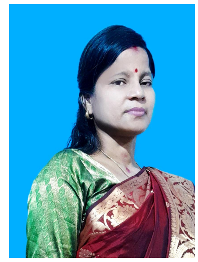
[25] S. M. Nayak and M. Rout, “A predictive model for bankruptcy: ANN, LSTM and CNN”, Journal of Statistics & Management Systems, Vol. 26, 2023, pp. 67-86.

[26] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique", Journal of artificial intelligence research, Vol. 16, 2002, pp. 321-357.

[27] S. Vellamcheti and P. Singh, "Class imbalance deep learning for bankruptcy prediction", First International Conference on Power, Control and Computing Technologies (ICPC2T), IEEE, 2020, pp. 421-425.

[28] H. J. Kim, N. O. Jo and K. S. Shin, "Optimization of cluster-based evolutionary undersampling for the artificial neural networks in

corporate bankruptcy prediction", Expert systems with applications, Vol. 59, 2016, pp. 226-234.
[29] D. S. Hong and C. Baik, "Generating and Validating Synthetic Training Data for Predicting Bankruptcy of Individual Businesses", Journal of Information & Communication Convergence Engineering, Vol. 19, No. 4 2021, pp. 228-238.
[30] H. Mansourifar, W. Shi, "Deep synthetic minority over-sampling technique", arXiv preprint arXiv:2003.09788, 2020.
[31] F. Mai, S. Tian, C. Lee, L. Ma, "Deep learning models for bankruptcy prediction using textual disclosures", European journal of operational research, Vol. 274, No. 2, 2019, pp. 743-758.
[32] N. Chandra, L. Ahuja, S. K. Khatri, H. Monga, "Utilizing Gated Recurrent Units to Retain Long Term Dependencies with Recurrent Neural Network in Text Classification", Journal of Information Systems and Telecommunication (JIST) , Vol.2, No.34, 2021, pp. 89-102.
[33] H. Kim, H. Cho, D. Ryu, "Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data", Computational Economics, Vol. 59, No. 3, 2022, pp. 1231-1249.
[34] H. Youn and Z. Gu, "Predict US restaurant firm failures: The artificial neural network model versus logistic regression model" Tourism and Hospitality Research, Vol. 10, No. 3, 2010, pp. 171-187.
[35] S. S. Park and M. Hancer, "A comparative study of logit and artificial neural networks in predicting bankruptcy in the hospitality industry", Tourism Economics, vol.18, No. 2, 2012, pp. 311-338.
[36] Y. Cao, X. Liu, J. Zhai, and S. Hua, “A two-stage Bayesian network model for corporate bankruptcy prediction”, International Journal of Finance & Economics, Vol. 27, No.1, 2022, pp. 455-472.
[37] S. Shetty, M. Musa, and X. Brédart, “Bankruptcy Prediction Using Machine Learning Techniques”, Journal of Risk and Financial Management, Vol. 15, No.1, 2022, pp.35.



Sasmita Manjari Nayak, is persuing her Ph.D. in Computer Science & Engineering, KIIT Deemed to be University.

Minakhi Rout, currently working as Associate Professor in school of computer engineering, KIIT Deemed to be University. She has recieved M.tech and Ph.D. degree in Computer Science & Engineering from Siksha ‘O’ Anusandhan University, Odisha, India in 2009 and 2015, respectively. She has more than 16 years of teaching and research experience in many reputed institute. Her research interest includes Computational Finance, Data Mining and Machine learning. She has published more than 50 research papers in various reputed journals and international conferences as well as guided several M.Tech and Ph.D. thesis. She is an editorial member of Turkish Journal of Forecasting.