A Novel Feature Selection Based Ensemble Approach to Bankruptcy Detection

MSc Research Project
MSc Data Analytics

Adebola Abdullahi-Attah
Student ID: X19119283

School of Computing
National College of Ireland

Supervisors: Dr Paul Stynes & Dr Pramod Pathak

MSc Project Submission Sheet
School of Computing
National College of Ireland

Student Name: Adebola Abdullahi-Attah ……. ……………………………………………………………………………………………

Student ID: X19119283 …………………………………………………………………………………………………..……

Programme: MSc Data Analytics ……………………………………………………………… **Year:** 2020 …………………………..

Module: MSc Research Project ………………………………………………………………………………………….………

Supervisor: Dr Paul Stynes & Dr Pramod Pathak ………………………………………………………………………………………….………

Submission Due Date: 17th August 2020 ………………………………………………………………………………………….………

Project Title: A Novel Feature Selection Based Ensemble Approach to Bankruptcy ………………………………………………………………………………………….………

Word Count: 7976 ……………………………………… **Page Count** 20 …………………………………………….……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: AAA ………

Date: 25TH August 2020 ………

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Novel Feature Selection Based Ensemble Approach to Bankruptcy Detection

Adebola Abdullahi-Attah
X19119283

Abstract

In domains such as finance, explainable machine learning approaches are important if they are to back decision-making systems especially in detecting potential bankruptcy where there are a myriad of attributes such as net profit, liquidity ratio that varies with each institution. With the business expansion and increase in company data, to foster explainable machine learning processes and potentially improve performance, there is a need to identify the most important indicators from highly dimensional data that can enhance bankruptcy detection and enable company owners to investigate financial statements with less need for external audit. Hence, this research investigates an optimal selection of features to detect bankruptcy using an ensemble approach combining six feature selection techniques namely Pearson's correlation, information gain, exhaustive feature selection, gradient boosting trees feature importance, random shuffling, and recursive feature elimination, through different voting mechanisms. For prediction, an ensemble approach is also employed combining random forest (RF), extreme gradient boosting (XGboost) and particle swarm optimized artificial neural network (PSO-ANN). The results indicate that ensemble approach for both feature selection and prediction outperformed state-of-the-art research in this domain with about 98% AUC score and 34 pertinent features were identified as major indicators of potential bankruptcy.

Keywords- bankruptcy, PSO-ANN, XGBoost, random forest, ensemble, financial statement.

1 Introduction

More recently, businesses have begun employing machine learning techniques to ascertain the financial state as regards potential bankruptcy but there is a need to present explainable models if these machine learning techniques are to be fully adopted to back decision-making processes. This challenge is further expressed in the number of business attributes that indicate the financial state of the business. Thus, to make the best model backed decisions on potential bankruptcy, care must be taken to select the best attributes for both ensuring optimum model performance and interpretation to stakeholders.

Outlining previous successes of machine learning techniques for predicting bankruptcy, Alaka et al. (2018) suggested that logistic regression and multiple discriminant analysis, support vector machine, artificial neural network, decision tree, genetic algorithm, rough sets, and case-based reasoning are the major techniques utilized but, identifying the best features that can adequately detect bankruptcy requires more research to enable seamless interpretation of financial statements and produce useable results. Hosaka (2019), Nyitrai and

Virág (2019) shared similar sentiments using random forest, support vector machine, and artificial neural network in their works to predict potential bankruptcy. Although insightful, they ignored feature selection processes that contribute to both model interpretability and performance.

The efficacy of machine learning models in predicting bankruptcy can be affected by the combination of different contributing variables. Therefore, experts devote adequate time to data preparation using different techniques to ensure data quality and integrity which is pivotal to making appropriate business decisions (Alaka et al., 2018). Different assumptions have been previously made by different researchers to select these attributes such as Kucher et al. (2018) that based his assumptions on the age of the company. This approach is however flawed as it is open to sample bias, exclusion bias, and missing values.

This research therefore proposes a novel ensemble approach for obtaining an optimal subset of features using statistical and classifier-based methods combined through different voting techniques and leveraging this subset to detect bankruptcy using ensemble classifiers. This thus poses the research question,

"To what extent can an ensemble approach to feature selection improve the detection of bankruptcy while identifying major indicators of potential bankruptcy from a financial statement?"

To carry out the research on Polish company bankruptcy data[1], subsequent research objectives were identified, and they are outlined as follows.

- Investigate recent research for the prediction of bankruptcy to ascertain best practices.
- Design an approach combining an array of statistical and classifier based for the features through hard majority voting (at least 4), a soft majority (at least 3) and minority voting (exactly 5), unanimous voting (all) and a novel any vote system (at least 1) for optimal selection of features
- Evaluate feature subsets obtained from the different voting systems using PSO-ANN
- Implement an ensemble of classifiers XGboost, random forest, and PSO-ANN
- Evaluate the approach using a comparative analysis between individual classifiers and different ensemble combinations using the area under curve and accuracy performance metrics.

The core contribution of this research to existing work on bankruptcy prediction is the novel rigorous approach employed to obtain an optimal feature subset that contributes to both model performance and interpretability to stakeholders. This novelty is also expressed in the ensemble approach adopted for prediction.

The rest of the paper is sectioned as follows. The related work section reflects previous pertinent researches that set the foundation and motivation for this current research. The methodology and design section present the experimental insight for the project, also the implementation section gives detailed steps in the construction of the specified models. The evaluation section shows the comparison and validation of the models implemented and the conclusion and future work section give the salient findings of the work with areas for future expansion on bankruptcy prediction.

[1] https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

2 Related Work

Over the years, insolvency detection research has been largely based on the discovery of algorithms that would improve models in terms of accuracy. The study of financial statements as a major means of improving existing research has therefore been investigated by Zieba et al. (2016) and all the features of the research have been implemented. Major algorithms used to investigate bankruptcy in the 1970s were based on discriminatory analysis and advanced to the logit and neural network in the 1980s and 1990s, all of these studies carried out predictions mainly using all the features, while some ignored the class imbalance in the data, which are some of the challenges of practical implementation (Alaka et al., 2018). The related literature review section provides a critical review of past articles and journals in this field, to justify the need for this research and contribute to the existing body of knowledge.

2.1 Background on Bankruptcy Prediction

Bankruptcy prediction have continued to be in the spotlight of research being a major determinant of nation's economic growth and financial viability for many stakeholders. Advancement in artificial intelligence has made it possible to improve the detection of bankruptcy and enhance intelligent decision making. Lyandres and Zhdanov (2013) investigated bankruptcy prediction based on theoretical and empirical analysis, the result showed that there is a strong positive relationship between investment opportunities and the probability of a company becoming bankrupt with an inference that companies with a high investment are less likely to renege in the commitment of paying their debts making them less susceptible to becoming bankrupt. Also, Uchida et al. (2015) shares a similar opinion with the above and applied the same methods of empirical analysis and added that investment of public funds into financial institutions increases their lending potentials thereby enhancing the value of the company. However, both pieces of researches identify explanatory parameters without taking cognizance of several changes in the company features thus, giving room for bias and irregular bankruptcy probability estimates which gave rise to discovering better techniques that utilized machine learning techniques for predictive analytics to improve the accuracy of results.

2.2 Machine Learning Approaches in Bankruptcy Prediction

Machine learning has been a major innovation in technology advancement, contributing to a better human life through a progressive imitation of human-based artificial intelligence by studying data, algorithms, and statistics to assist people make smart business decisions. Thus, Son et al. (2019) proposed a machine learning approach that can enhance the accuracy of bankruptcy prediction results through thorough data processing and the implementation of boosting gradient for the selection of features of which the proposed model improved on existing research with an AUC of 17% increase over existing approach reviewed. However, this work was limited due to the fact that selecting features based on a single technique hinders optimal performance due to information loss (Nalic et al., 2020), this challenge has been addressed in this research by the implementation of six feature techniques. Also, Mai et al. (2019) carried out a study of bankruptcy using a combination of text and numeric data, while applying a deep learning approach of convolutional neural network CNN, random forest, and support vector machine. The experiments were carried out without investigation of class balance in the numeric data. The results also showed low scores in the evaluation metrics of AUC in the experiments which could have been caused by imbalanced data.

Further research undertaken by Jan (2018), Uthayakumar et al. (2020) in bankruptcy and financial risk prediction considered several decision machine learning algorithms such as artificial neural network (ANN), support vector machine (SVM), ant colony optimization (ACO-FCP). The research recorded a high accuracy of about 91% and 98% respectively without investigating the class imbalance in the data before conducting the experiments using a similar dataset of high dimensions for the prediction. Although the experiments showed very high accuracy yet for this nature of data, accuracy may not be the valid evaluation metrics. This hypothesis is expressed in the research conducted by Zou et al. (2016), which discourages the use of accuracy as a suitable metrics for an imbalanced data and discloses accuracy as a biased metric for such task due to evaluating the higher class and not both classes thus making the area under ROC a more suitable evaluative metrics for such imbalance data. To improve on existing research, further experiments were conducted by Song and Peng (2019) to study class imbalance in financial data prior to modelling using multiple sampling techniques such as SMOTE, boarder line SMOTE and random sampling and the implementation of AUC, F1 score to measure model performance. Investigating this approach improved the model performance with a minimum AUC score of 2%.

2.3 Machine Learning and Feature Engineering Approaches in Bankruptcy Prediction

Subsequent increase in data dimension have driven the need for feature engineering techniques to tackle feature redundancy and enhance explainable machine learning approaches using several feature selection techniques based on filter, wrapper, and embedded approaches. Filter and wrapper based methods have their individual limitations such as the filter method been limited in recognizing the relationship between variables and the wrapper may cause overfitting thus, the application of both techniques forming a hybrid feature selection technique.

Exploring the recent heuristics wrapper-based method, Uthayakumar et al. (2020) investigated four heuristics-based feature selection techniques namely genetic algorithm (GA), particle swarm optimization (PSO), gray wolf optimization compared to the proposed ant colony optimization (ACO) using five datasets. On each dataset, the methods are implemented singly to justify the output of each function algorithm. Measuring the efficiency of the features selected by each of them, ACO-FS selected the highest number of features on the Polish data set 32 features at a cost of 0.02, GA selected 30 at a cost of 0.25, PSO selected 18 attributes at a cost of 0.75 and GWO selected 5 at a cost of 0.04. The overall result shows ACO as the best feature selector with the largest number of features at the lowest cost. The high performance of the algorithm made it ideal for classification and even outperformed other algorithms, such as the artificial neural network. Similar research was done by Son et al. (2019), contributed to existing research by finding feature importance for models that aided the higher performance of the models. The prediction was done using XGboost, LightGBM, and other algorithms and measured with the metrics of AUC. Xgboost gave the overall best result of 83% on all features and 88% AUC with selected features, this shows that selection of features aided the model performance. Lin et al. (2019) researched efficient feature selection and classification techniques using the waikato environment knowledge analysis (WEKA) framework. The genetic algorithm and principal component analysis (PCA) were used for selecting features and the wrapper based genetic algorithm outperformed the filter method of PCA. Also, naïve bayes and support vector machine are used as classifiers for the model in which the naïve bayes error value reduced from 8 percent to 4.97 percent. This shows the essence of feature selection in reducing the rate of error and improve the performance of machine learning models.

2.4 Ensemble Approaches to Bankruptcy prediction

The ensemble approach in machine learning comprises of a combination of weak learners to enable an optimal predictive performance above the application of the single learners. Due to the continued research on improving the existing work in bankruptcy detection, the ensemble approach has shown to outperform single learners (Tsai et al., 2014).

An overview study using ensemble approaches in the financial sector was done by Tsai et al. (2014) and outline favourable machine learning ensemble methods as bagging, boosting and stacking. Majorly the boosting approach has been largely applied in predictive analytics using boosting algorithms such as the extreme gradient boosting (XGBoost) which is a combination of several decision trees, adaptive boosting (Adaboost) comprises of single weak learners. The boosting approach of extreme gradient boosting is applied in the research by Zieba et al. (2016) which functions based on boosted decision trees and randomly selects new features. The performance of the models is measured using the area under roc and 10-fold cross-validation and the results of the experiments across the five years of the dataset used to show the ensemble approach to produce high scores of 90% and above. Similarly, Chen et al. (2020) applied the label proportions as ensemble learning. The approach combined both the boosting and bagging with the support vector machine (SVM). The proportion SVM was distributed into K-groups of bags and then stacked using the predictions of the multiple proportions SVMs. Hence applying the bagging and stacking ensembled approach for the task. The results also showed the bagging ensembled approach outperformed the boosting method by having the lowest error rate.

In conclusion, as can be seen from reviewed research feature engineering is a very important aspect of the machine learning process as it directly impacts the performance of the algorithm and offers some level of explainability especially in recent times where the use of machine learning models for algorithm decision making has been called into question due to their black-box approaches. This is especially the case in financial endeavours like the prediction of bankruptcy. Although this has been partially investigated for bankruptcy in reviewed research, there does not appear to be research that has adopted an ensemble voting approach combining classifier, statistical, and evolutionary algorithm-based methods. As ensemble methods have been shown to improve predictive accuracy, adopting a similar approach in carefully preparing the features for the algorithms will make for even better predictive performances and produce explainable models.

3 Research Methodology

This section fully outlines the procedures employed to carry out the research. These are divided into steps data collection and description, data processing, feature engineering and modelling procedures employed to carry out the research experiment. Figure 1 illustrates the methodology process flow.

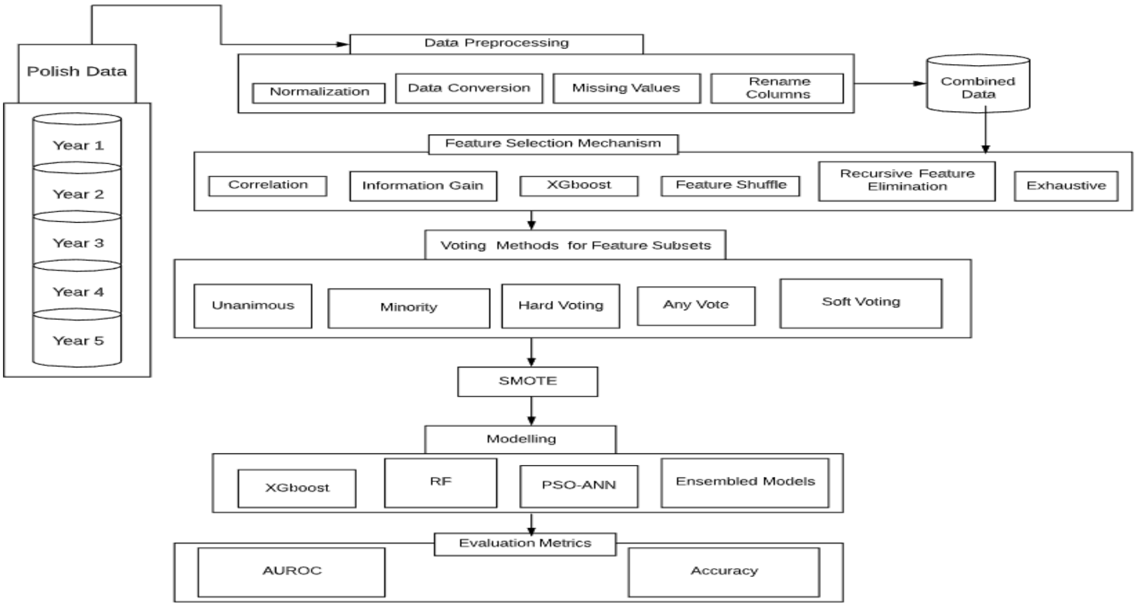


Figure 1: Methodology process flow

3.1 Data Collection and Description

The pedagogical research for this study is proposed by Son et al. (2019). To ascertain the performance of the proposed approach, the polish companies’ bankruptcy dataset (Zieba et al., 2016) is used. The multivariate dataset shows the bankruptcy status of companies across five years and consists of 64 features such as net profit/total assets, profit on operating activities/financial expenses retained earnings/total assets with 43,405 instances comprising of numeric values and a predicted class of bankruptcy and non-bankruptcy. The dataset was obtained from the University of California Irvine (UCI) data repository and is publicly available for research purposes. Further data cleaning, preparations and transformation is done and are outlined in the next sections.

3.2 Data Preparation

The quality of the data is essential to the output of the results. Thus, there is need to prepare the data rigorously before modelling. For this research, the Polish data has several missing values and some attributes that are not essential for modelling. The data type is converted to float and normalized across all attributes in the list. The missing values are imputed using a multivariate imputation using a chained equation (MICE) that considers the uncertainties in all variables while replacing the missing values (Barnes and Palas, 2019). These were done on the annual datasets before combination to a single dataset of 64 features and 43,405 instances (Uthayakumar et al., 2018).

3.3 Feature Selection

To select optimum number of features, six feature engineering techniques of pearson correlation, information gain, gradient boosting, recursive feature elimination, exhaustive search was carried out to construct the voting mechanism. This was further used to generate the feature subsets of unanimous votes (including all techniques), any vote (at least selected by one technique), hard voting (at least four techniques selected), soft voting (at least three methods selected) and minority voting (exactly five of the techniques selected). The correlation technique identified 31 uncorrelated features, thereby setting the selection of best 31 features from other selection methods (Nalic et al., 2020). Also, the synthetic minority oversampling technique (SMOTE) was applied to resolve the class imbalance in the data which is a ratio of 2091: 41,264.

3.4 Modelling algorithms

3.4.1 Optimized Artificial Neural Network (ANN) using Particle swarm optimization (PSO).

The PSO-ANN algorithm is implemented on the four feature subsets from the feature engineering phase for both balanced and imbalanced data. The algorithm is also ensembled with RF and XGboost algorithms to form PSO-ANN/RF and PSO-ANN/XGboost ensembles aiming to improve performance. Figure 2 below shows the architecture of the ANN with 8 features.

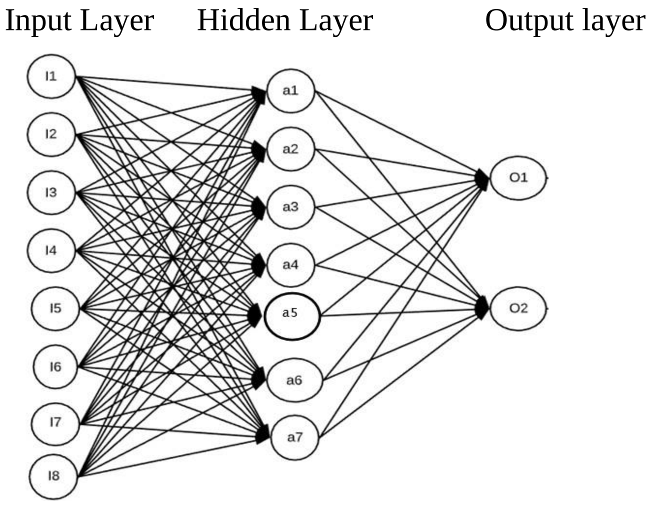


Figure 2: Neural network architecture for minority feature subset.

3.4.2 Random Forest (RF)

The RF classifier is applied on the four feature subsets following the construction of the feature engineering phase for both balanced and unbalanced data. The algorithm is also combined with SVC and XGboost algorithms for the ensembled approach for the purpose of enhancing performance.

3.4.3 Extreme Gradient Boosting (XGBoost)

The XGboost algorithm is implemented on the four feature subsets from the feature engineering phase for both balanced and unbalanced data. The algorithm is also combined with SVC and RF algorithms to construct SVC_RF_XGboost with the purpose of improving performance.

3.4.4 Support Vector Classifier (SVC)

The support vector machine is implemented to achieve the ensemble approach of XGboost and RF.

3.5 Evaluation Metrics

The classification task of bankruptcy prediction is class-oriented, and this supported the selection of metrics that can adequately evaluate the models with cognizance of their class disparity. The metrics of accuracy may lead to misleading measures because it evaluates based on the class with the higher observations especially the negative class in binary classification. Hence, the suitability of area under the receiver operating characteristics (AUROC) for this research which measures the extent to which the models can identify the difference amongst the classes. This also functions based on the true positive rate (TPR) and the false positive rate (FPR) using the threshold framework (Son et al., 2019). The metrics of accuracy were also explored.

4 Design Specification

4.1 Feature Selection Algorithms

To solve the bias and limitation of using a single selection approach. Therefore, six robust feature selection techniques have been applied individually and then combined through a voting mechanism to create four subsets of features that would be used as inputs into the models.

- Pearson correlation: This indicates the measure of linear correlation between two variables. (Wang et al., 2019).
- Information gain: This measures the gain of each variable with respect to the dependent variable. (Alhaj et al., 2016).
- XGboost: This selects optimal features using the estimations from a trained prediction model to rank features in the order of importance (Son et al., 2019).
- Recursive feature elimination: This performs feature selection in a backward medium thereby dropping features with the least predictive power (Hou 2020).
- Feature shuffle: The technique outputs features that are uncorrelated with the dependent variable using mean (Naik and Mohan, 2019).
- Exhaustive feature selection: This examines all possible feature combinations while selecting the features with the most informative characteristics for the prediction. (Mnich and Rudnicki, 2020).

4.2 Voting Mechanism

The voting technique enables the creation of four different ensembled feature subsets based on their frequency counts using four conditions of minority voting (exactly 5), Hard voting (at least 4), soft voting (at least 3), any vote (at least 1) and unanimous (all) selections on the feature selection techniques (Nalic et al., 2020), The feature subsets of hard voting and soft voting have same feature similar features. Thus, the soft voting feature subset was ignored in the implementation.

4.3 Optimized Artificial Neural Network (ANN) using Particle swarm optimization (PSO).

ANN is developed to mimic the processing pattern of the human brain using neurons represented as nodes which are employed in modelling and decision making for a complex task such as face recognition and credit risk. The types of neural network consist of the feedforward, kohonen self-organizing neural network, recurrent neural network. The learning process of each node is based on the signal received from other nodes and processes it while assigning weights using the transfer function to give the calculated output. The network architecture consists of the input, hidden, and output layers. However, some of the limitations faced by ordinary artificial neural networks include computational costs and large variations between inputs and outputs, and the PSO is therefore, adopted to improve the reliability of the model (Qu et al., 2019).

PSO is an evolutionary optimization technique that is patterned after the behaviour of social animals such as birds flow and fish school. This function using several particles known as a swarm. The search is carried out using a set of the randomly generated population of swarms with each in the swarm termed as a particle. The learning process is based on the environment and individual experience (Bansal, 2019).

4.4 Random Forest (RF)

RF is an ensembled robust tree-based classifier that utilizes the bagging approach of randomly selecting subsets using the attribute of the nodes to avoid overfitting in the data. It then gives the output from the most votes in the subsets to enable high accuracy in the prediction (Barboza et al., 2017).

4.5 Extreme Gradient Boosting (XGboost)

XGboost is an ensemble classifier that is based on the boosting approach which gives it a strong advantage above the base classifiers. It frequently adds up weak classifiers in sequence to improve the output of the model and stops iteration when there is no better performance. This enables it to ensure higher accuracy and minimized errors in the predictive performance of the model (Son et al., 2019).

4.6 Support Vector Classifier (SVC)

The support vector machine is an efficient modelling technique that accommodates a variety of tasks such as regression and classification. The algorithm has high learning capacity with few features, computationally efficient and durable with a challenged dataset (Song and Peng, 2019).

5 Implementation

The experiments were carried out using the python programming language in Jupyter notebook in Google Colab environment. The table 1 below outlines the tools and the implementation process flow has been demonstrated in figure 3.

Table 1: A table of implementation Process

S/No	Technologies for the research implementation	Purpose
1	Microsoft Excel	For storage and retrieval of the Polish data in CSV format
2	Google Colab	The Working environment consists of python libraries.
3	Jupyter Notebook	For implementing python codes
4	Python	For executing machine learning task using high-level codes

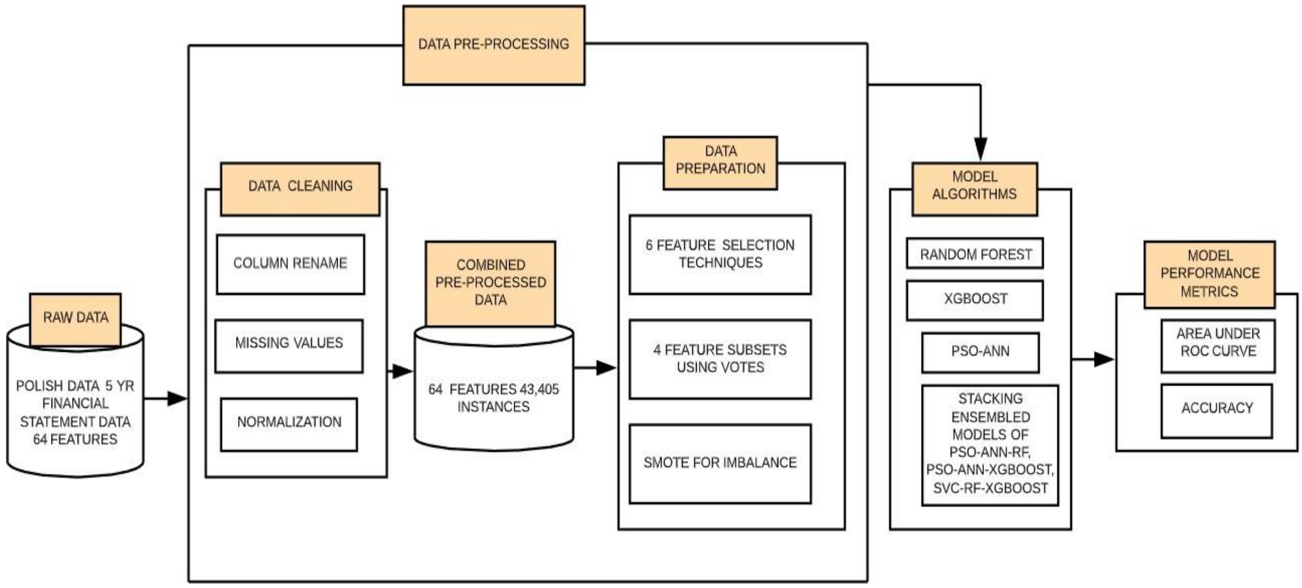
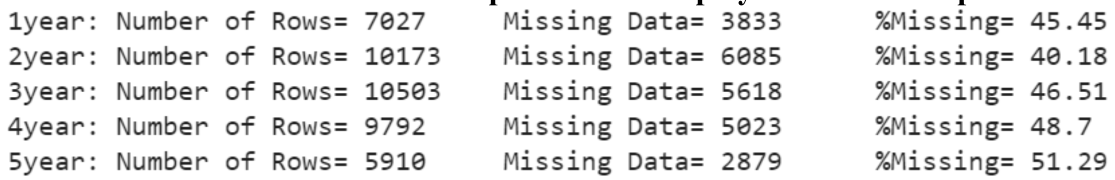

DATA PRE-PROCESSING
RAW DATA
POLISH DATA 5 YR FINANCIAL STATEMENT DATA 64 FEATURES
DATA CLEANING
COLUMN RENAME
MISSING VALUES
NORMALIZATION
COMBINED PRE-PROCESSED DATA
64 FEATURES 43,405 INSTANCES
DATA PREPARATION
6 FEATURE SELECTION TECHNIQUES
4 FEATURE SUBSETS USING VOTES
SMOTE FOR IMBALANCE
MODEL ALGORITHMS
RANDOM FOREST
XGBOOST
PSO-ANN
STACKING ENSEMBLED MODELS OF PSO-ANN-RF, PSO-ANN-XGBOOST, SVC-RF-XGBOOST
MODEL PERFORMANCE METRICS
AREA UNDER ROC CURVE
ACCURACY

Figure 3: Implementation Process Flow

5.1 Data Handling

The dataset is extracted in CSV format as arff data files. The dataset has 64 features such as net profit/total assets and 43,405 instances altogether. The python libraries of pandas, NumPy, and Scipy enabled the reading of the data into data frames and further used for the exploratory data analysis. Table 2 gives the description of the dataset.

Table 2: **A table on Polish companies' bankruptcy data set description**



```
1year: Number of Rows= 7027     Missing Data= 3833      %Missing= 45.45
2year: Number of Rows= 10173    Missing Data= 6085      %Missing= 40.18
3year: Number of Rows= 10503    Missing Data= 5618      %Missing= 46.51
4year: Number of Rows= 9792     Missing Data= 5023      %Missing= 48.7
5year: Number of Rows= 5910     Missing Data= 2879      %Missing= 51.29
```

The columns in the data which was represented as attr were changed to the meta-data description of X1 to X64 to enhance visualizations and statistical description of the data. Missing values were investigated using the sparsity matrix which showed the measure of missing values was large enough to affect the quality of the output if dropped. Furthermore, normalization of the data was carried out and MICE was applied to impute the missing values using the fancyimpute python library. The five-year data is combined into a single dataset for feature selection and modeling.

The features are selected using six feature selection techniques namely Pearson's correlation, information gain, exhaustive feature selection, gradient boosting trees feature importance, random shuffling, and recursive feature elimination, through different voting mechanisms which was applied individually and the output of best 31 features from each of the selection method is then combined through a voting mechanism to create four subsets of minority voting (exactly 5), Hard voting (at least 4), any vote (at least 1) and unanimous (all) features that would be fed as inputs into the models. Table 3 represents the voting techniques.

Table 3: A table of voting mechanism for feature subsets

Total Features	Unanimous selection (all 6)	Minority (Exactly 5 selection)	Hard Voting (At least 4selection)	Soft Voting (At least 3 selection)	Any Vote (At least 1 selection)
64	20	8	31	31	34

The class imbalance in the dataset was investigated in which the majority class consists of the non-bankrupt class of 0's and the synthetic majority oversampling techniques (SMOTE) was used to reduce the class imbalance in the data.

5.2 Prediction Models

The Extreme gradient boosting and random forest have been implemented as single classifiers and as ensembled approach with the PSO-ANN and SVC. The number of estimators was set as 200 for both algorithms while the maximum tree depth in XGboost is set as 10 and radial kernel was used for SVC. The models are implemented using scikit learn and is also used as part of the ensemble approach.

The PSO enables the ANN to achieve minimal loss by optimizing the weights and bias parameters of the ANN. The architecture of the ANN is built using the feedforward propagation with the function f to calculate the loss values. The input layer corresponds to the number of input features, while the hidden layer is calculated based on the rule of thumb of 2/3 the size of the input layer, with the addition of the output layer. Also, the output layer is based on predictive classes. The dimensions for the shape of input-to- hidden weight matrix is based on the number of inputs and the hidden layer, the number of hidden layers gives the shape of input-to- hidden bias array, the shape of the hidden-to-output weight matrix is calculated using the number of hidden layer and outputs, lastly, the shape of hidden -to-output bias array is known using the output value. Hence, unrolling them together gives the dimensions for each particle in the swarm. The parameters implemented for the optimization are c1=0.5, c2=0.3, w=0.9, no of iterations are 1000. The total number of particles used is 100 and for each of the iteration, the loss cost is calculated and applied for optimizing the values of the particles. The pyswarms python package enabled the implementation of PSO optimization.

6 Results and Evaluation

To evaluate the performance of the models the widely used metrics accuracy and AUC have been applied because the AUC can effectively differentiate between the classes especially when using an imbalanced data and to enable comparison with the base paper. Table 4 explains all six bankruptcy prediction experiments involving XGboost, RF, PSO-ANN, PSO-ANN-RF, PSO-ANN-XGboost, SVC-RF-XGboost across all four feature subsets of unanimous, minority, hard voting and any vote. Research by Son et al. (2019) is also used as a basis for comparison. Experiments with good results in both imbalance and balanced data have been highlighted, while the best overall results are identified in red ink in the single and ensemble models.

Table 4: **A table on model performance results**

AUC Score Based on a Single Feature Selection Approach of XGboost

	XGboost	88%
	random forest	88%
Son et al., 2019	ANN	85%

Proposed Approach using Six Feature Selection Techniques

Models	AUROC				Accuracy			
Non-Smote	Un-animous	Minority	Hard Voting	Any Vote	Un-animous	Minority	Hard Voting	Any Vote
PSO-ANN	0.5605	0.5018	0.5822	0.5070	0.9510	0.9508	0.9506	0.9510
RF	0.5842	0.5104	0.5760	0.5717	0.9574	0.9496	0.9569	0.9562
XGBoost	0.6242	0.5110	0.6275	0.6240	0.9582	0.9462	0.9570	0.9579
PSO-ANN-RF	0.6004	0.6132	0.5619	0.5376	0.9634	0.9572	0.9482	0.9528
PSO-ANN-XGB	0.5943	0.6093	0.6629	0.6280	0.9629	0.9582	0.9629	0.9735
SVC-RF-XGBOOST	0.8491	-	0.8401	0.8401	0.9599	-	0.9595	0.9595
Smote								
PSO-ANN	0.9656	0.9431	0.9751	0.9699	0.9680	0.9512	0.9751	0.9699
RF	0.9695	0.9294	0.9750	0.9762	0.9694	0.9293	0.9750	0.9762
XGBoost	0.9697	0.9093	0.9772	0.9795	0.9696	0.9092	0.9772	0.9794
PSO-ANN-RF	0.9709	0.9398	0.9197	0.9669	0.9727	0.9442	0.9366	0.9647
PSO-ANN-XGB	0.9682	0.9426	0.9767	0.9689	0.9777	0.9372	0.9769	0.9715
SVC-RF-XGBOOST	0.9715	0.9298	0.9784	0.9797	0.9715	0.9298	0.9784	0.9797

Experiment 1: XGBoost Models with four feature subsets

In the first experiment, rigorous pre-processing was done in terms of imputting missing values and selection of features to prepare the data for the experiment in which four different subsets of minority voting, hard voting, any vote and unanimous voting were created. The results in table 4 showed that XGboost obtained the highest AUC score of a minimum of 62% across three of the four subsets created in the imbalance data. This is an indication of the ability of XGboost to handle imbalance data. Further applying the algorithm of XGboost on balanced data about 98% AUC were gotten from the feature subset of any vote consisting of 34 features. This shows that the subsets of 34 features can give a good performance of the model and improve on existing research of Son et al. (2019). Thus, outlining the importance of carefully curated feature selection approaches such as the ensemble approach used.

6.1 Experiment 2: RF Models with the Feature Subsets

In the second experiment, the random forest model achieved an AUC score of less than 60% across all the feature subsets with the imbalance data while on the balanced data it is considered to be a good classifier because all the results, as seen in table 4 improved from 0.50 to above 0.90. This shows the importance of solving class imbalance to improve the predictive performance of a classification model. The overall best performance of RF model is given by any voting feature subset that further demonstrates that any vote feature subset can enhance model performance while indicating that these features are relevant in the prediction of bankruptcy.

6.2 Experiment 3: PSO-ANN Models

With an AUC score of 0.975 the PSO improved the performance of the ANN model in comparison with Son et al. (2019). The model performed well in that it gave a high AUC score of 0.975 with the feature subset of hard voting but relatively less performance when compared to the general performance of the bagging and boosting algorithms. This result validates the hypothesis of ensemble algorithms to outperform single learners even when used as individual classifiers. The PSO reduced overfitting in the ANN model which is a major challenge for ANN models.

6.3 Experiment 4: Ensembled Algorithms of PSO-ANN-XGB & PSO-ANN-RF

The results in table 4 show the approach of the ensemble to perform well. The PSO-ANN-XGboost having AUC scores of 0.9767 with hard voting subset outperformed the PSO-ANN-RF with 0.9709 AUC with unanimous voting subset. This gives an insight into the inference that, when optimal features are selected, the state-of-the-art optimized improved ANN makes less contribution to improving the performance of ensemble algorithms as single classifiers. However, this model can be improved by increasing the number of features and adjusting the parameters used for the ANN architecture.

6.4 Experiment 5: Ensembled Algorithms of SVC-XGBoost-RF

The results obtained from the ensemble model in table 4 illustrates the performance of the stacking approach of support vector machine with XGboost and RF over the PSO-ANN ensembled approach of boosting and bagging algorithms applied in the task. The model gave the overall best performance in the prediction of bankruptcy with an AUC score of 0.9797. Also, the feature subset of any vote of 34 features enhanced the model performance. Furthermore, the confusion matrix for the ensembled model in table 5 shows the subset of unanimous features to give the lowest scores of 178 representing misclassified bankrupt companies as non-bankrupt and 171 misclassified non-bankrupt companies as bankrupt, giving the lowest error rate. However, the model was limited in giving the results for the imbalance data in the subset of minority which may be due to the small size of input features.

The results further validate the assumption of ensembled approach outperforming single learning approach even when they are ensemble single classifiers.

Table 5: A table on confusion matrix for ensembled approach of SVC-RF-XGBoost using balanced data

	Predicted True Positive				**Predicted False Positive**			
Actual Positives	Minority	Unanimous	Hard Voting	Any Vote	Minority	Unanimous	Hard Voting	Any Vote
	8112	8434	8511	8538	621	221	178	178
Actual Negatives	**Predicted False Negatives**				**Predicted True Negatives**			
	595	273	196	171	8008	8408	8451	8449

These results indicate that an ensemble feature selection approach improves the detection of bankruptcy while also identifying major indicators of potential bankruptcy thus, answering the research question.

6.5 Discussion

In this study, four feature subsets have been obtained from an ensemble feature selection technique using the voting mechanism of unanimous, minority, hard voting, and any vote to identify the best subsets for the prediction of bankruptcy. The first experiment conducted is using XGboost algorithm as single classifiers and this is considered the baseline experiment, also other algorithms of PSO-ANN and random forest have been implemented as single learners and ensemble. The results of all the experiments have been recorded in table 5 with high performing models and feature subsets highlighted in the balanced and imbalanced data. The red font colour shows the overall best-performing models as single and ensemble learners.

The performance of the models' results showed that adequate pre-processing in terms of filling missing values and the ensemble feature selection enhanced the achievement of XGBoost resulting in high AUC scores on the balanced data. In comparison with Son et. al. (2019) the approach enabled a substantial increase of 0.8 AUC over the existing results. The adoption of a PSO training strategy for the ANN led to a better performance compared to the base paper. Furthermore, the random forest model also had significant improvement over the existing research. The ensemble feature subset of any vote of 34 features provides better performance over the other subsets and the feature subsets of hard voting with 31 features demonstrated to be a good indicator of bankruptcy because it was well represented by 8 of the models making it a more prominent feature indicator of bankruptcy. The results obtained based on the feature subset of minority which consist of 8 features support the hypothesis that implementing all features for machine learning task is better to avoid information loss. However, in research such as this, in which identification of feature is essential to the decision making, how features are reduced or dropped should be carefully done to achieve optimal performance of prediction models since the approach gave high results of 90% AUC and yet limited the optimal performance of the models. The overall performance of XGboost as a single classifier and in the ensemble approach outperformed other models built making it highly recommended for the financial prediction task and the ensemble approach to feature selection is proposed over single approach when using a high dimensional data. Although the difference in the AUC scores of the models are minimal and may seem statistically insignificant, however, ignoring such values is increasing the error rate and contributing to human bias in the predictive task. The bar plots in figure 4,5,6,7 show the performance of the feature subsets of unanimous, minority, hard voting, and any vote, respectively. The feature subsets were constructed from an ensemble approach using four different voting mechanism of all (unanimous), exactly five (minority), at least four (hard voting) and at least one (any vote) features selected from the application of six feature selection techniques of Pearson correlation, feature shuffle, extreme gradient boosting feature importance, exhaustive search, information gain and recursive feature elimination. These feature subsets were used as inputs features for the models of extreme gradient boosting (XGboost), random forest (RF), particle swarm optimized artificial neural network (PSO-ANN), and the ensembled models of PSO-ANN-RF, PSO-ANN-XGboost and SVC-RF-XGboost. The performance of these models was measured using the AUC and accuracy metrics.

Unanimous feature subset

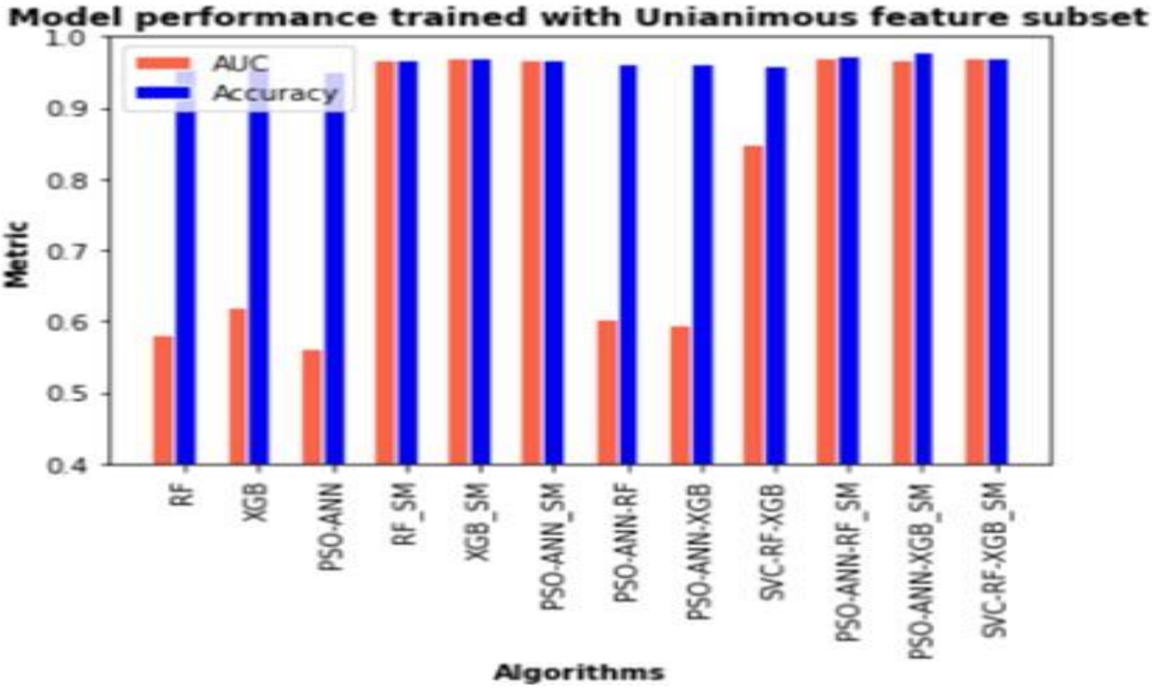


Figure 4: Unanimous Subset

The performance of the unanimous feature subset is shown in figure 4, the subset of which consist of 20 features that have been selected by all six feature selection techniques applied. These features are then used as inputs for XGboost, RF, PSO-ANN, SVC-XGboost-RF, PSO- ANN-RF, and PSO-ANN-XGboost models. Experiments have been carried out on balanced and imbalanced data, and their performance is measured using the area under curve and accuracy. In all six experiments, the lowest AUC score was 0.56 in the imbalanced data of the PSO-ANN model and the highest AUC score was

0.9715 from the balanced data of the ensembled approach of xgboost and random forest.

Minority feature subset

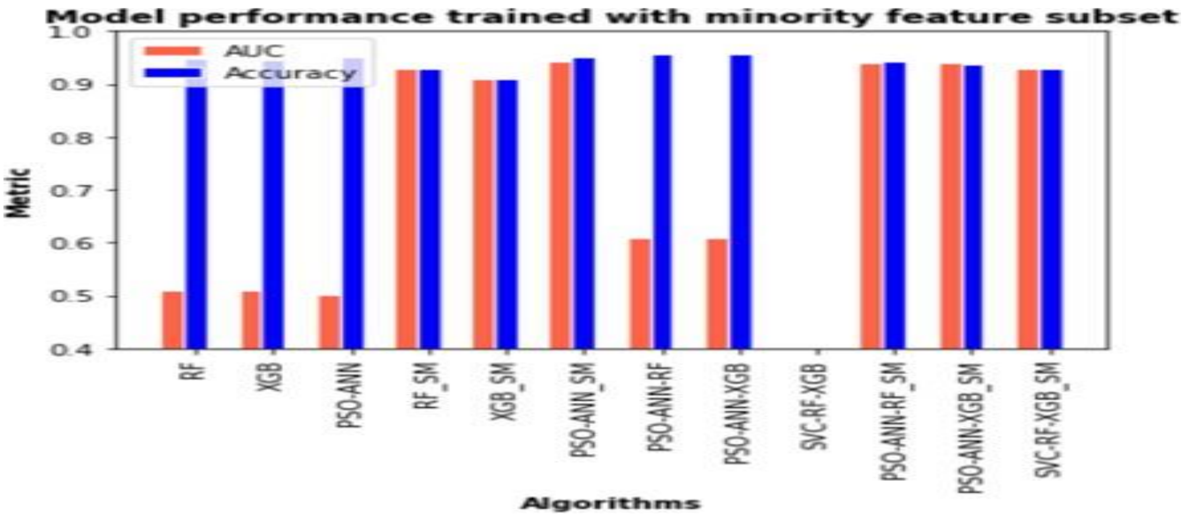


Figure 5: Minority Subset

The feature subset of the minority voting mechanism consist of 8 features that have been selected based on the number of counts that have shown that five of the feature selection techniques selected these features. This provides the lowest number of inputs for the models. Figure 5 shows the performance of the subset in all the six experiments conducted; the feature subset provided the lowest performance with a maximum AUC score of 0.9426 from the ensembled approach of PSO-ANN-XGboost. This gives an insight into the impact of input features on the overall performance of the decision-making classification model, given that the approach yielded high results of 94% AUC and yet limited the optimal performance of the decision-making process. It was also limited in the output performance of the imbalance data in the ensemble approach of XGboost and random forest.

Hard voting feature subset.

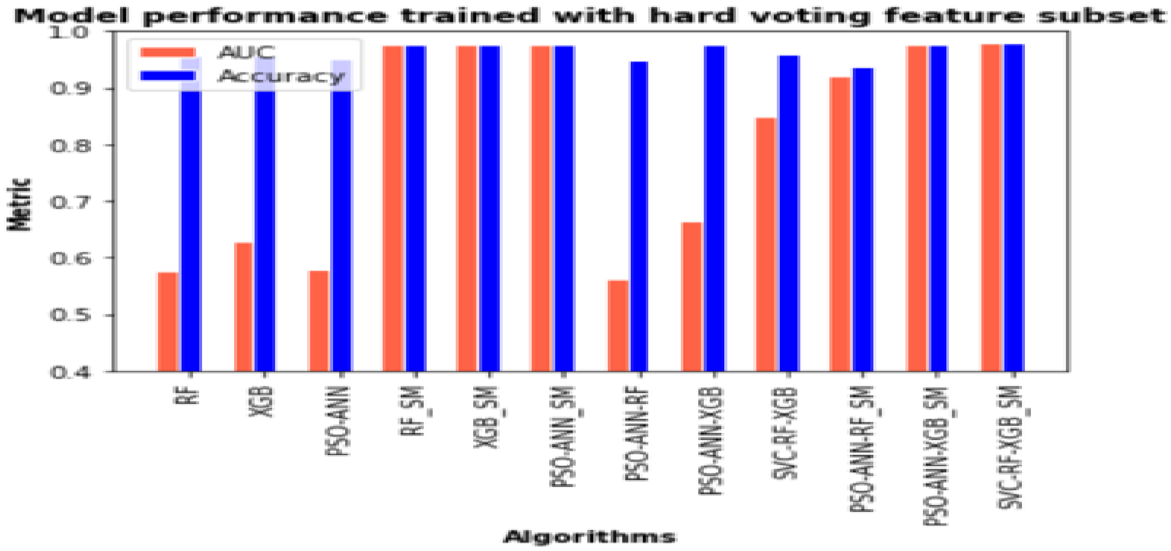


Figure 6: Hard voting feature subset

Figure 6 shows the overall performance of the hard-voting feature subset consisting of 31 features that have been constructed using the voting mechanism that at least 4 of the six feature selection techniques applied have chosen those features. The 31 features are applied as inputs to the six models and the results of the models show that this resulted in higher AUC scores over the models built with minority and unanimous feature subsets.

Anyvote feature subset

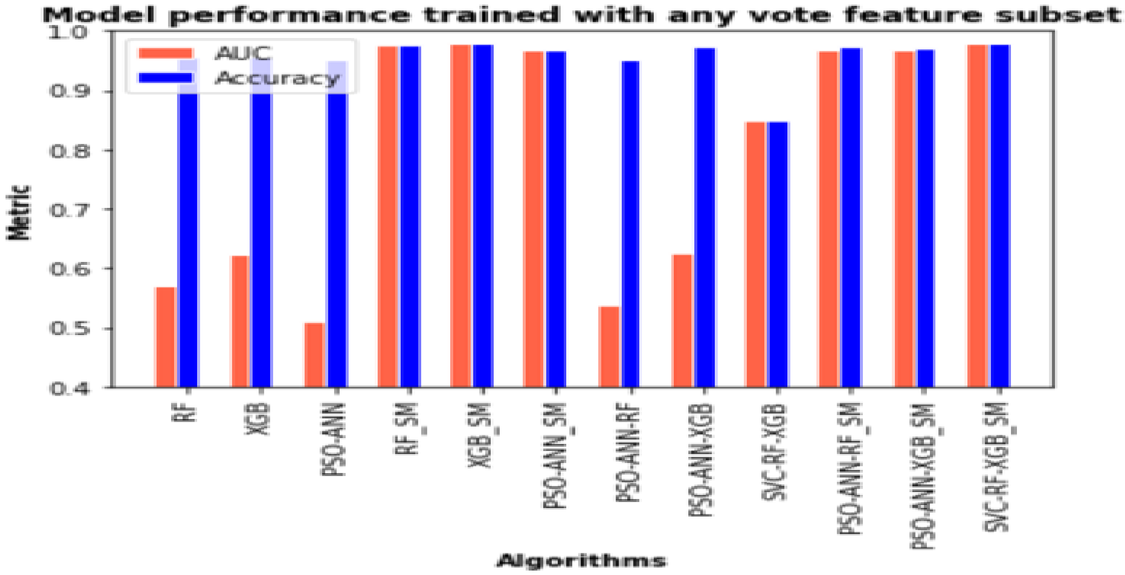


Figure 7: Any vote feature subset

The feature subset of any vote in figure 7 was constructed using the voting mechanism if any of the six feature selection techniques used in the research selected the feature, resulting in a total of 34 features that were further used as inputs for the six models. This approach gave the overall best results on the balanced data with a maximum AUC score of 0.9795 for the individual classifier and 0.9797 for the ensemble model of XGboost and RF. This reinforces the essence of careful selection of features to achieve optimal model performance and accurate decision- making.

7. Conclusion and Future Work

Machine learning techniques have greatly advanced forecasting ability in many fields of endeavour such as the financial sector. Identifying pertinent features driving this forecasting and improving explainability is a challenge in this domain. Thus, this work investigates to what extent can an ensemble feature selection approach improve the detection of bankruptcy while identifying major indicators of potential bankruptcy. As part of the objectives to accomplish this research, six feature selection techniques namely Pearson’s correlation, information gain, exhaustive feature selection, gradient boosting trees feature importance, random shuffling and recursive feature elimination were ensembled through different voting mechanisms. Each ensemble voting technique resulted in a different subset of features for classification. From the results of the experiments, it is evident that the ensemble feature subset of any vote significantly improved the performance of the XGboost models when applied singly and in an ensemble giving the highest AUC score of 97.9%.

The results from the confusion matrix with same subset of any vote gave the lowest error rate this implies that the single approach to feature selection limits machine learning algorithms in achieving optimal decision making process and increases the level of error in the research. Furthermore, it is observed that the XGboost algorithm attained the highest AUC scores in three of four feature subsets of unanimous, hard voting and any vote making it an efficient and suitable algorithm for financial task. Also, the ensemble approach to bankruptcy prediction outperformed the single learning which supports the existing hypothesis of ensemble methods producing higher performance over the single learners. Overall, all the experiments gave good performance above existing researches in this domain. Further hyperparameter optimization could have further improved the performance of the approaches but was not done due to computational resource implications. For future work, the proposed approach is recommended to be implemented on high dimensional data.

8. Acknowledgement

I am most thankful to God for helping me to complete this program and for his mercies during the length of the course. I am also indebted to my incredible supervisors Dr Paul Stynes and Dr Pramod Pathak for their unwavering support, guidance, and commitment which has been a source of encouragement throughout the research project. Finally, I would like to express my sincere appreciation to all my lecturers, coursemates and family for their tremendous support throughout the course.

References

Alaka, H.A., Oyedele, L.O., Owolabi, H.A., Kumar, V., Ajayi, S.O., Akinade, O.O. and Bilal, M. (2018) ‘Systematic review of bankruptcy prediction models: Towards a framework for tool selection’, *Expert Systems with Applications*, 94, pp. 164-184, ScienceDirect. doi: 10.1016/j.eswa.2017.10.040

Alhaj, T.A., Siraj, M.M., Zainal, A., Elshoush, H.T. and Elhaj, F. (2016) ‘Feature selection using information gain for improved structural-based alert correlation’. *PloS one*, 11(11), p.e0166017. doi: 10.1371/journal.pone.0166017

Bansal, J.C. (2019) ‘Particle swarm optimization’, *In Evolutionary and swarm intelligence algorithms*, pp. 11-23, Springer, Cham. SpringerLink. doi: 10.1007/978-3-319-91341-4_2.

Baranes, A. and Palas, R. (2019) ‘Earning Movement Prediction Using Machine Learning-Support Vector Machines’, *Journal of Management Information and Decision Sciences*, 22(2), pp. 36-53, ProQuest.

Barboza, F., Kimura, H. and Altman, E. (2017) ‘Machine learning models and bankruptcy prediction’. *Expert Systems with Applications*, 83, pp. 405-417, ScienceDirect doi: 10.1016/j.eswa.2017.04.006

Chen, Z., Chen, W. and Shi, Y., 2020. ‘Ensemble learning with label proportions for bankruptcy prediction’, *Expert Systems with Applications, 146,* p.113155. ScienceDirect. doi: 10.1016/j.eswa.2019.113155

Hosaka, T. (2019) ‘Bankruptcy prediction using imaged financial ratios and convolutional neural network’, *Expert systems with Applications,* 117, pp. 287-299, ScienceDirect. doi: 10.1016/j.eswa.2018.09.039.

Hou, X. (2020) ‘P2P borrower default identification and prediction based on rfe-multiple classification models’, *Open Journal of Business and Management,* 8(2), pp. 866-880, Scientific Research. doi: 10.4236/ojbm.2020.82053

Jan, C.L. (2018) ‘An effective financial statements fraud detection model for the sustainable development of financial markets: Evidence from Taiwan’, *Sustainability,* 10(2), p. 513, MDPI. doi: 10.3390/su10020513.

Kücher, A., Mayr, S., Mitter, C., Duller, C. and Feldbauer-Durstmüller, B. (2018) ‘Firm age dynamics and causes of corporate bankruptcy: Age dependent explanations for business failure’, *Review of Managerial Science,* pp. 1-29, SpringerLink. doi: 10.1007/s11846- 0180303-2.

Lin, W.C., Lu, Y.H. and Tsai, C.F. (2019) ‘Feature selection in single and ensemble learning based bankruptcy prediction models’, *Expert Systems,* 36(1), p.e 12335, Wiley Online Library. doi: 10.1111/exsy.12335

Lyandres, E. and Zhdanov, A. (2013) ‘Investment opportunities and bankruptcy prediction’, *Journal of Financial Markets*, 16(3), pp. 439-476, ScienceDirect. doi: 10.1016/j.finmar.2012.10.003

Mai, F., Tian, S., Lee, C. and Ma, L. (2019) ‘Deep learning models for bankruptcy prediction using textual disclosures’, *European Journal of Operational Research*, 274(2), pp. 743-758, ScienceDirect. doi: 10.1016/j.ejor.2018.10.024

Mnich, K. and Rudnicki, W.R. (2020) ‘All-relevant feature selection using multidimensional filters with exhaustive search’, *Information Sciences,* 524, pp. 277-297, ScienceDirect. doi: 10.1016/j.ins.2020.03.024

Naik, N. and Mohan, B.R. (2019) ‘Optimal feature selection of technical indicator and stock prediction using machine learning technique’, in 2019 *International Conference on Emerging Technologies in Computer Engineering (ICETCE).* Springer, Singapore, 1 February 2019, pp. 261-268, SpringerLink. doi: 10.1007/978-981-13-8300-7_22.

Nalić, J., Martinović, G. and Žagar, D. (2020) ‘New hybrid data mining model for credit scoring based on feature selection algorithm and ensemble classifiers’, *Advanced Engineering Informatics,* 45, p. 101130, ScienceDirect. doi: 10.1016/j.aei.2020.101130

Nyitrai, T. and Virág, M. (2019) 'The effects of handling outliers on the performance of bankruptcy prediction models', *Socio-Economic Planning Sciences*, 67, pp. 34-42, ScienceDirect. doi: 10.1016/j.seps.2018.08.004.

Qu, Y., Quan, P., Lei, M. and Shi, Y. (2019) 'Review of bankruptcy prediction using machine learning and deep learning techniques', *Procedia Computer Science*, *162*, pp. 895- 899, ScienceDirect. doi: 10.1016/j.procs.2019.12.065.

Son, H., Hyun, C., Phan, D. and Hwang, H.J. (2019) 'Data analytic approach for bankruptcy prediction', *Expert Systems with Applications*, 138, p. 112816, ScienceDirect. doi: 10.1016/j.eswa.2019.07.033

Song, Y. and Peng, Y. (2019) 'A MCDM-based evaluation approach for imbalanced classification methods in financial risk prediction', *IEEE Access*, 7, pp. 84897-84906. doi: 10.1109/ACCESS.2019.2924923

Tsai, C.F., Hsu, Y.F. and Yen, D.C. (2014) 'A comparative study of classifier ensembles for bankruptcy prediction', *Applied Soft Computing*, 24, pp. 977-984. ScienceDirect. doi: 10.1016/j.asoc.2014.08.047

Uchida, H., Miyakawa, D., Hosono, K., Ono, A., Uchino, T. and Uesugi, I. (2015) 'Financial shocks, bankruptcy, and natural selection', *Japan and the World Economy*, 36, pp. 123-135, ScienceDirect. doi: 10.1016/j.japwor.2015.11.002

Uthayakumar, J., Metawa, N., Shankar, K. and Lakshmanaprabu, S.K. (2020) 'Financial crisis prediction model using ant colony optimization', *International Journal of Information Management*, 50, pp. 538-556, ScienceDirect. doi: 10.1016/j.ijinfomgt.2018.12.001.

Uthayakumar, J., Metawa, N., Shankar, K. and Lakshmanaprabu, S.K. (2018) 'Intelligent hybrid model for financial crisis prediction using machine learning techniques', *Information Systems and e-Business Management*, pp. 1-29, SpringerLink. doi: 10.1007/s10257-018- 0388-9

Wang, L., Nie, C. and Wang, S. (2019) 'A new credit spread to predict economic activities in China'. *Journal of Systems Science and Complexity*, 32(4), pp. 1140-1166, SpringerLink. doi: 10.1007/s11424-019-8033-3

Zięba, M., Tomczak, S.K. and Tomczak, J.M. (2016) 'Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction', *Expert Systems with Applications*, 58, pp. 93-101, ScienceDirect. doi: 10.1016/j.eswa.2016.04.001

Zou, Q., Xie, S., Lin, Z., Wu, M. and Ju, Y. (2016) 'Finding the best classification threshold in imbalanced classification', *Big Data Research*, 5, pp. 2-8, ScienceDirect. doi: 10.1016/j.bdr.2015.12.001