Improving Prediction Accuracy using Random Forest Algorithm

Nesma Elsayed[1*], Sherif Abd Elaleem[2], Mohamed Marie[3]
Business Information Systems Department-Faculty of Commerce and Business Administration,
Helwan University, Helwan, Egypt[1]
Business Administration Department-Faculty of Commerce and Business Administration, Helwan University, Helwan, Egypt[2]
Information Systems Department-Faculty of Computers and Artificial Intelligence, Helwan University, Helwan, Egypt[3]

Abstract*—One of the latest studies in predicting bankruptcy is the performance of the financial prediction models. Although several models have been developed, they often do not achieve high performance, especially when using an imbalanced data set. This highlights the need for more exact prediction models. This paper examines the application as well as the benefits of machine learning with the purpose of constructing prediction models in the field of corporate financial performance. There is a lack of scientific research related to the effects of using random forest algorithms in attribute selection and prediction process for enhancing financial prediction. This paper tests various feature selection methods along with different prediction models to fill the gap. The study used a quantitative approach to develop and propose a business failure model. The approach involved analyzing and preprocessing a large dataset of bankrupt and non-bankrupt enterprises. The performance of the model was then evaluated using various metrics such as accuracy, precision, and recall. Findings from the present study show that random forest is recommended as the best model to predict corporate bankruptcy. Moreover, findings write down that the proper use of attribute selection methods helps to enhance the prediction precision of the proposed models. The use of random forest algorithm in feature selection and prediction can produce more exact and more reliable results in predicting bankruptcy. The study proves the potential of machine learning techniques to enhance financial performance.

***Keywords*—Corporate bankruptcy; feature selection; financial ratios; prediction models; random forest**

I. Introduction

Predictions in business are essential tools for decision-making and strategic planning. At its core, a prediction is an educated guess about what the future holds based on past trends and current data. When used correctly, predictions can help businesses prepare for various scenarios and make informed decisions.

It is a common fact that there is no certainty in the field of business. Prediction models can provide decision makers with a framework to set more realistic strategies via predicting financial performance. In the case of predicting a business failure, management can prevent business bankruptcy. Bankruptcy prediction helps in increasing accuracy of decision making process for business enterprises since it has a variety of applications in financial fields [1].

The key idea is that public information of corporations comprises significant data and information that could be used by investors to asses financial status, which may be a major reason to cause bankruptcy [2]. Financial crisis prediction indicators included Profitability, Solvency, Growth ability, Cash flow and Capital structure [3]. Enhanced prediction accuracy is bound to increase the earnings to shareholders by improving financial risk management inside rising markets [4].

Recent research has employed financial ratios to show the exploration models for business failure. To improve prediction accuracy, it is important to find the most influential factors on financial performance. The discriminatory influence acquired by bringing together distinctive groups of financial ratios (FRs) and corporate governance indicators (CGIs) for business failure prediction was examined [5].

It is worth mentioning that the massive amount of corporate data presents an opportunity to deeply analyze the data and, consecutively, gain a great deal of knowledge. Unfortunately, the necessity for many human resources and too much time limit the benefits of the financial data. Alternatively, improving machine learning techniques can save both time and money. This helps to provide decision makers with significant evidence to be a base for making strategic plans.

In past works, a variety of prediction models were applied to define the early warning factors of a potential bankruptcy. This paper attempts to examine and compare the significance of using decision tree, k-nearest neighbor, logistic regression, multilayer Perceptron, and random forest in predicting corporate failure.

In 2022 a study used only three financial indicators: the return on assets, the current ratio, and the solvency ratio reported prediction accuracy rates of more than 80 percent. The study used Belgian companies` data set contains a sample of 3728 Belgian companies that were announced bankrupt between 2002 and 2012 to anticipate bankruptcy [6].

The main research gap is that “the performance of prediction models attained by combination of various categories of FRs has not been completely investigated. Only some chosen FRs have been utilized in previous researches and the selected attributes may vary from study to study [7].

The goal of our study is using random forest algorithm for analyzing corporate data encompassing various goals. Firstly, it aims to evaluate the tendency of business failure in different companies by developing prediction models that incorporate random forest algorithms in both attribute selection process and prediction process. Secondly, the model eases the enhancement of prediction process by enabling researchers to foresee the influence of fluctuations in ninety-five different financial ratios on corporate financial performance. Additionally, the research contributes by developing a novel model applying random forest algorithm along with seven various categories of financial indicators to anticipate business failures.

Paper structure consists of literature review in Section II providing a clear explanation of previous studies in bankruptcy prediction field, methods in Section III presenting our model that is based on incorporating the most common algorithms in financial prediction field, results in Section IV demonstrating the average performance measures for prediction models used in our study, discussion section clarifying the importance of our model and the significance of our contribution.

II. Literature Review

A. Corporate Bankruptcy Prediction

The substantial rise in the total papers, especially subsequent the 2008 global financial disaster, has verified that corporate bankruptcy is a subject of growing interest, which indicates the importance of this issue for corporations [8]. Regrettably, the COVID-19 epidemic that has invaded the globe since 2020 was one of the major triggers for bankruptcy filing. While data analytics has many applications in the financial field, Bankruptcy Prediction Models (BPM) have witnessed an increase in recognition [9].

Beaver assessed various financial variables to evaluate their ability in classifying and predicting bankrupt firms. That made him a pioneer in researches that study the enterprise failure prediction [10]. Altman presented business failure models according to discriminant study in categorizing economic failure based on five financial ratios: working capital/total assets, market value of equity/total debt, earnings before interest and taxes/total assets, retained earnings/total assets and sales/total assets [11].

Once Altman issued one of the very popular models in prediction of firms bankruptcy in 1968, a variety of models that predict bankruptcy have been issued in the literature [12]. It does not only direct attention to the increasing number of research issued, but also to the diversity of enterprise failure prediction models employed for business crisis prediction. Owing to the advance in machine learning methods and computer ability in latest years, more diverse analytical tools have been employed to create a business failure model with superior precision.

B. Prediction Accuracy Enhancement

There are two steps to assess financial crisis. Whereas the first stage employs a variety of financial variables, the other one employs diverse classifiers in construction of the bankruptcy model [13].

First, in the financial variables step, selecting the most informative financial variables can improve prediction performance. Regarding to Chih-Fong Tsai investigational outcomes, applying attribute selection tools to choose and extract the extra valuable, demonstrative and illustrative variables can reduce the effort and time of training the model, which certainly results in increasing the performance of prediction [14].

Second, in the model construction step, a variety of methods have been offered, including decision tree, k-nearest neighbor, logistic regression, multilayer Perceptron, and random forest. In 1980, Ohlson estimated the probabilities of bankruptcy employing logistic regression [15].

One of the early applications of random forests was reported in 2001 that presented random sampling of trees and the concept of tree correlation. His discoveries indicated that for the first-time forest algorithms can rival with arcing approaches, in both classification and regression analysis [16]. Separate research done in 2012 showed that RF is effective for more accurate results. This assists the researchers in estimating feature significance and value [17].

Artificial neural networks (ANNs) are powerful artificial intelligence technologies that are widely-used as they are able to combine several nonlinear functions to express non-linear relationships between input data and a class label [18]. A previous study on corporate distress prediction examined the precision of Logit and ANN to establish a comparison between using statistical and artificial intelligence in modeling financial risk [4].

III. Methods

A. Dataset

The data set is acquired from (UCIMLR)[19], which supplies datasets to the interested researchers in machine learning field. Initially, the sample data was gathered by the Taiwan Economic Journal. It comprises the financial variables of industrial, electronic, shipping, tourism, and retail companies for the years 1999–2009. The data set includes ninety-five various financial indicators and 6,819 rows, of which 6,599 are corporations that did not go bankrupt, and 220 are bankrupt corporations. The description of business failure is established on the rules of the Taiwan Stock Exchange. Our proposed model is shown in Fig. 1.

B. Preprocessing

In In data preprocessing stage, we checked for missing values and duplicates within the dataset, but there were none. We applied data normalization on the dataset. All preprocessing and feature selection steps were conducted using WEKA.

An imbalanced data set is created when the total observations in one group exceeds the total of observations in the other group. Prediction techniques behave disappointingly in data sets with imbalanced classes because they regularly suppose that all classes are represented equally.

As a result, the cases that are represented in the smaller group are miscategorized as belonging to the mass group [20].

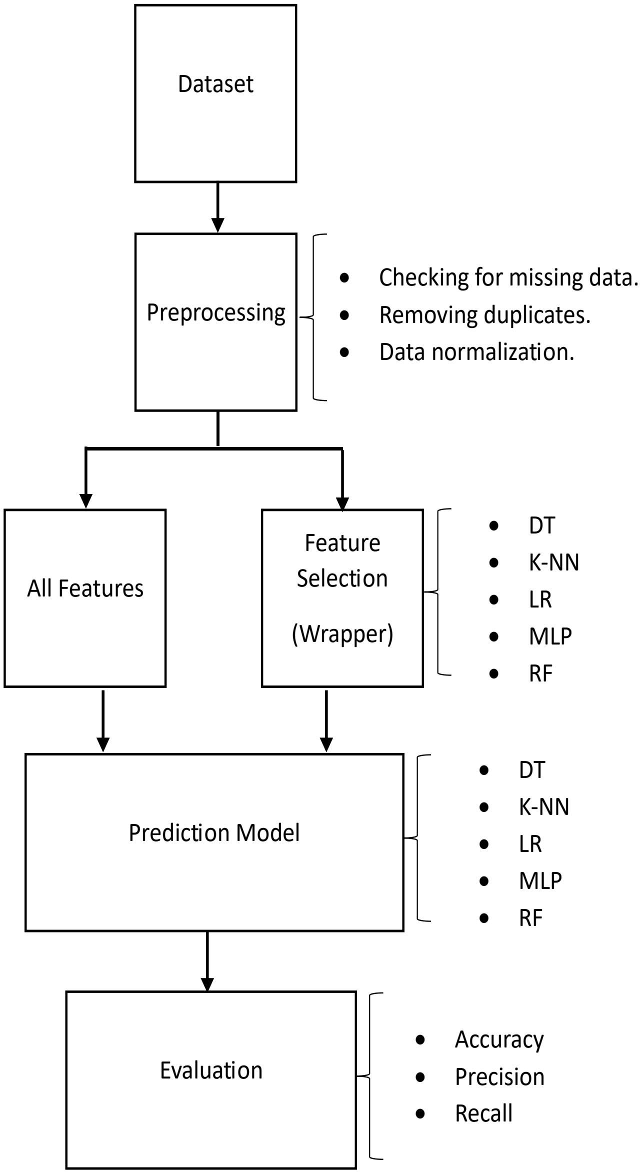

Fig. 1 Our proposed model.

In each year, the total of business failures is smaller in comparison with the total of companies that did not go bankrupt. If failed corporations are outliers, this causes a key breach to the fundamental distributional conjectures for logistic regression [21]. Resampling techniques generate new samples of data from the original dataset using a set of statistical methods. It is essential to lower the danger of the study or machine learning algorithm biasing toward the common class.

We applied unsupervised resample filter on data to get more reliable results by producing a random subsample of a dataset. It applies over sampling on the smaller group and under sampling on the mass group at the same time while keeping the same number of records in the original dataset. Thus, using unsupervised resampling helps in gaining the benefits of both over and under-sampling. That leads to having more reliable and realistic results.

C. Feature Selection

Attribute selection is the practice of selecting the important attributes that have an influence on the performance of the model. Attribute selection is a research problem in wrapper methodology, so different combinations are made, assessed, and compared with other combinations. The algorithm is trained by using the subset of features iteratively.

In the present study, the wrapper method is applied on the resampled data since it interacts with classifier, models feature dependencies, minimizes computational cost, and provides good classification accuracy.

Features chosen will differ regarding the kind of classifier as diverse classifiers perform better with various arrays of attributes to generate more competent conclusions. The five classifiers which will be employed for the feature electing manner outcomes are illustrated in Table I.

TABLE. I. FEATURES SELECTED USING WRAPPER METHOD

Classifier	Attributes selected based on the wrapper method
Decision Tree (DT)	X8, X10, X12, X40, X55, X64, X65, X87, X92
K-nearest Neighbor (KNN)	X9, X14, X21, X25, X31, X52, X54, X62, X73, X85, X87, X90, X92
Logistic Regression (LR)	X11, X13, X17, X18, X21, X27, X34, X39, X44, X51, X64
Multilayer Perceptron (MLP)	X2, X3, X16, X32, X39, X43, X49, X50, X52, X59, X61, X68, X73, X76, X84
Random Forest (RF)	X34, X40, X48, X50, X54, X68, X76, X77, X80, X90, X91, X93

The highest significant features for effective bankruptcy prediction are the categories of solvency and profitability [5]. All classifiers have selected attributes from both categories except random forest algorithm has not selected any attributes from solvency category. RF algorithm has selected more attributes from growth category than other classifiers.

D. Prediction

To recognize the best bankruptcy model, different techniques have been applied on the data set, and then their outcomes are matched with each other. Models are established according to two distinctive situations:

All attributes will be employed, and only the features chosen using the wrapper method will be employed.

We trained models utilizing the same five algorithms employed in features selection wrapper method to analyze the relation between employing the same algorithm in both attribute selection stage and prediction modeling stage.

E. Evaluation

In cross-validation the data set is indiscriminately divided into 'k' groups. Only one group is treated as a test set whereas the extra groups are treated as training sets. The training sets aim to teach the model while the test set is utilized to assess the model. The activity is done repeatedly until every distinctive group has been utilized as the test set.

Cross-validation test is preferred to be used in such cases because it offers the model the chance to learn on multiple train-test divisions. This provides us with a well sign of how accurate the model will operate on undetected data.

In this research, the following metrics are employed to estimate model performance: accuracy, precision, and recall. The metrics computation is established on rules presented in Table II. Computation of accuracy, precision, recall and F-measure are constructed on confusion matrix involving a classification of actual and predicted values into the next groups: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

Accuracy demonstrates how often a machine learning model is correct overall. Depending only on accuracy measure to estimate model performance can be deceiving when utilizing imbalanced data set as it assigns equal weight to the classes which mitigate the model's capability to predict all classes. Precision presents how frequently a machine learning model is precise when predicting the intended category. Recall expresses whether a machine learning model can recognize all objects of the intended category.

TABLE. II. PERFORMANCE METRICS

Evaluation measure	Rule
Accuracy	$\frac{TP + TN}{Total}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$

IV. RESULTS

Since bankruptcy is an imbalanced problem, then the weighted average is preferred for measuring performance of classification models. Table III in Appendix summarizes the scores of the evaluation process. Considering these results, we can state that some models such as KNN, MLP, and RF work better with the features chosen utilizing wrapper method with the same algorithm. On the contrary, some models such as decision tree and logistic provided better results employing all attributes.

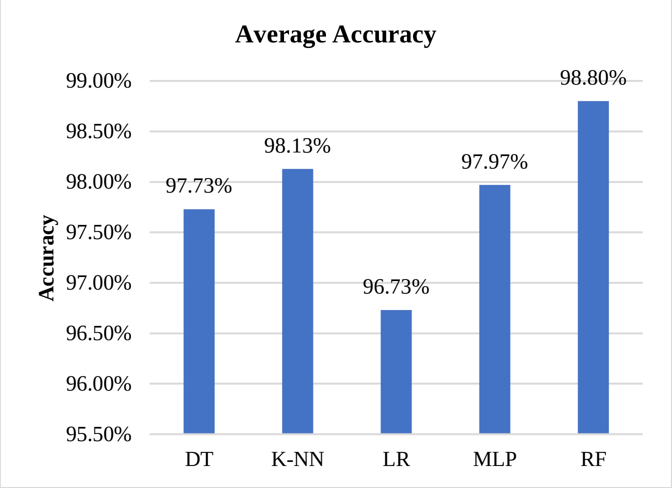

Fig. 2 Average accuracy of each model.

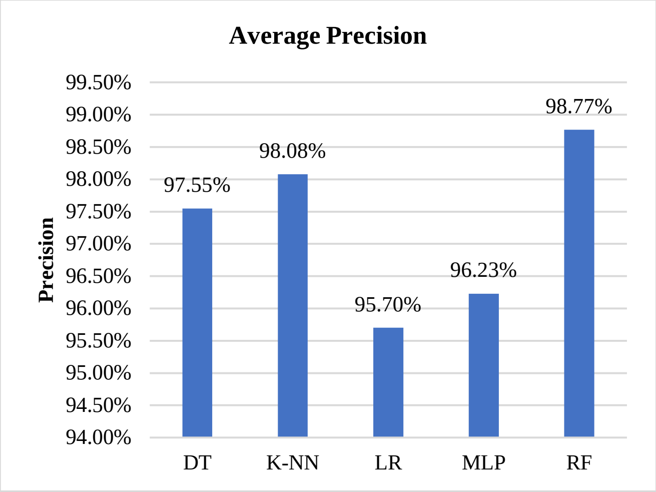

Fig. 3 Average precision of each model.

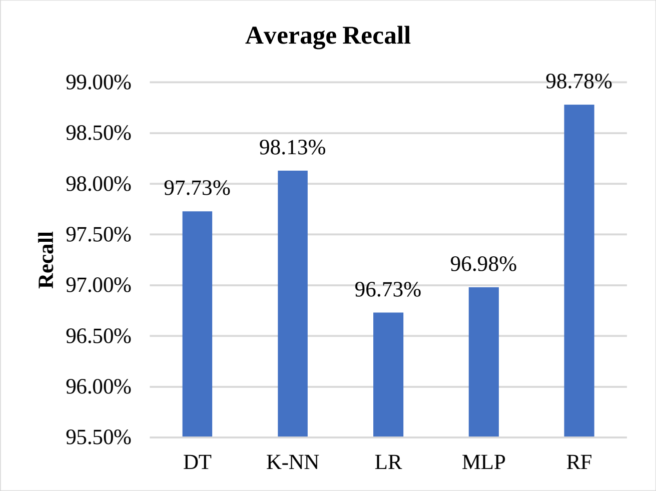

Fig. 4 Average recall of each model.

In analyzing Fig. 2, we see that the RF is better than the other models concerning illustrating and comparing the differences in average accuracy of models. As shown in this figure, the random forest was the most accurate model with 98.80% compared with the lowest accuracy of 96.73% for logistic model.

Fig. 3 displays the average precision of models employing different feature selection algorithms. Once more, the data suggests that RF, on average, provide an improved performance than their counterparts. This finding also sides with the preceding idea that RF is more efficient. While random forest was the most precise model with 98.77% the logistic model with 95.70% was the lowest in precision.

Fig. 4 also proves that random forest model performs better than other models. Again, we notice the same tendencies of RF model exceeding other models. Random forest model was the most model able to correctly identify most of the positive results with 98.78% sensitivity while the logistic model only had 96.73% which was the lowest. We can state that the model using random forest technique outperformed all other models in all performance metrics in predicting bankruptcy.

V. Discussion

To further confirm our findings, we compared them to other studies using the same sample dataset of Taiwanese enterprises along with various resampling, attributes selection and prediction techniques.

In 2016, the models published by [5] employed Support Vector Machine (SVM) and generated five different machine learning models. Along with the ninety-five financial ratios we utilized, they also used CGIs. They employed 10-fold cross-validation to generate ten distinct training and test samples. They also tried five alternative attribute selection techniques. The model with the best performance in their study achieved 81.5% accuracy that was exceeded by the weakest model in our research.

In 2022, the research by [22] closely examined the discriminatory competence of a MLP in studying financial failure prediction. For this purpose, they employed different setups of optimization algorithms, activation functions, number of neurons, and number of layers. The model with the best performance in their study achieved 86.67% accuracy, 95.47% precision and 85.24% sensitivity that was outperformed by the worst model in our study.

VI. Conclusions

This research covers the usage of different techniques with the aim of enhancing the findings of prediction. We can state that firstly, using feature selection can significantly improve performance of prediction models. Secondly, constructing prediction models using random forest algorithms outperformed other models using different machine learning techniques in terms of accuracy, precision, and sensitivity. Thirdly, employing growth ratios in dataset used in financial failure prediction is significant. Results from this study recommend that, in general, random forest algorithms tend to attain more exact results. The impressive performance of the random forest model can be improved when the wrapper method is used as the attribute selection method with random forest algorithm to detect the best features for the classifier. Practitioners can benefit from these conclusions to enhance the accuracy of their predictions. For future work, researchers may use different feature selection methods combined with a diversity of resampling approaches to identify what works better.

Funding: None.

Conflicts of Interest: None.

References

[1] Y. Zhang et al., "Towards augmented kernel extreme learning models for bankruptcy prediction: algorithmic behavior and comprehensive analysis," Neurocomputing, vol. 430, pp. 185-212, 2021, doi: http://dx.doi.org/10.1016/j.neucom.2020.10.038.

[2] Z. Huang, H. Chen, C.-J. Hsu, W.-H. Chen, and S. Wu, "Credit rating analysis with support vector machines and neural networks: a market comparative study," Decision support systems, vol. 37, no. 4, pp. 543-558, 2004, doi: http://dx.doi.org/10.1016/S0167-9236(03)00086-1.

[3] M. Jiang and X. Wang, "Research on intelligent prediction method of financial crisis of listed enterprises based on Random Forest algorithm," Security and Communication Networks, vol. 2021, pp. 1-7, 2021.

[4] L. Muparuri and V. Gumbo, "On logit and artificial neural networks in corporate distress modelling for Zimbabwe listed corporates," Sustainability Analytics and Modeling, vol. 2, p. 100006, 2022, doi: http://dx.doi.org/10.1016/j.samod.2022.100006.

[5] D. Liang, C.-C. Lu, C.-F. Tsai, and G.-A. Shih, "Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study," European journal of operational research, vol. 252, no. 2, pp. 561-572, 2016.

[6] S. Shetty, M. Musa, and X. Brédart, "Bankruptcy Prediction Using Machine Learning Techniques," Journal of Risk and Financial Management, vol. 15, no. 1, p. 35, 2022.

[7] D. Liang, C.-F. Tsai, H.-Y. R. Lu, and L.-S. Chang, "Combining corporate governance indicators with stacking ensembles for financial distress prediction," Journal of Business Research, vol. 120, pp. 137-146, 2020.

[8] Y. Shi and X. Li, "An overview of bankruptcy prediction models for corporate firms: A systematic literature review," Intangible Capital, vol. 15, no. 2, pp. 114-127, 2019, doi: http://dx.doi.org/10.3926/ic.1354.

[9] S. C. Mann and R. Logeswaran, "Data Analytics in Improved Bankruptcy Prediction with Industrial Risk," in 2021 14th International Conference on Developments in eSystems Engineering (DeSE), 2021: IEEE, pp. 23-26, doi: http://dx.doi.org/10.1109/DeSE54285.2021.9719372.

[10] W. H. Beaver, "Financial Ratios As Predictors of Failure," Journal of Accounting Research, vol. 4, pp. 71-111, 1966, doi: 10.2307/2490171.

[11] E. I. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," The Journal of Finance, vol. 23, no. 4, pp. 589-609, 1968, doi: 10.1111/j.1540-6261.1968.tb00843.x.

[12] J. L. Bellovary, D. E. Giacomino, and M. D. Akers, "A review of bankruptcy prediction studies: 1930 to present," Journal of Financial education, pp. 1-42, 2007.

[13] F. Lin, D. Liang, and E. Chen, "Financial ratio selection for business crisis prediction," Expert systems with applications, vol. 38, no. 12, pp. 15094-15102, 2011, doi: http://dx.doi.org/10.1016/j.eswa.2011.05.035.

[14] C.-F. Tsai, "Feature selection in bankruptcy prediction," Knowledge-Based Systems, vol. 22, no. 2, pp. 120-127, 2009, doi: http://dx.doi.org/10.1016/j.knosys.2008.08.002.

[15] J. A. Ohlson, "Financial ratios and the probabilistic prediction of bankruptcy," Journal of accounting research, pp. 109-131, 1980, doi: http://dx.doi.org/10.2307/2490395.

[16] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001, doi: https://doi.org/10.1023/A:1010933404324.

[17] D. Sharma, "Improving the art, craft and science of economic credit risk scorecards using random forests: Why credit scorers and economists should use random forests," Craft and Science of Economic Credit Risk Scorecards Using Random Forests: Why Credit Scorers and Economists Should Use Random Forests (June 9, 2011), 2011, doi: http://dx.doi.org/10.2139/ssrn.1861535.

[18] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618," Genetic programming and evolvable machines, vol. 19, no. 1-2, pp. 305-307, 2018, doi: http://dx.doi.org/10.1007/s10710-017-9314-z.

[19] Taiwanese Bankruptcy Prediction, doi: https://doi.org/10.24432/C5004D.

[20] Y. F. Roumani, J. K. Nwankpa, and M. Tanniru, "Predicting firm failure in the software industry," Artificial Intelligence Review, vol. 53, pp. 4161-4182, 2020, doi: http://dx.doi.org/10.1007/s10462-019-09789-2.

[21] R. P. Hauser and D. Booth, "Predicting bankruptcy with robust logistic regression," Journal of Data Science, vol. 9, no. 4, pp. 565-584, 2011, doi: http://dx.doi.org/10.6339/JDS.201110_09(4).0006.

[22] R. F. Brenes, A. Johannssen, and N. Chukhrova, "An intelligent bankruptcy prediction model using a multilayer perceptron," Intelligent Systems with Applications, p. 200136, 2022.

APPENDIX

TABLE. III. COMPARISON BETWEEN THE OUTCOMES OF USING DIFFERENT ATTRIBUTE SELECTION AND MODELING ALGORITHMS

Prediction model	Feature selection	Accuracy	Precision	Recall
DT	None	98.0056%	.979	.980
DT	DT Wrapper	98.0496%	.979	.980
DT	KNN Wrapper	97.6536%	.974	.977
DT	LR Wrapper	97.375%	.971	.974
DT	MLP Wrapper	97.8003%	.976	.978
DT	RF Wrapper	97.4776%	.974	.975
KNN	None	98.1816%	.981	.982
KNN	DT Wrapper	98.1522%	.981	.982
KNN	KNN Wrapper	98.5922%	.985	.986
KNN	LR Wrapper	97.8003%	.978	.978
KNN	MLP Wrapper	97.9909%	.980	.980
KNN	RF Wrapper	98.0496%	.980	.980
LR	None	97.1257%	.967	.971
LR	DT Wrapper	96.5244%	.949	.965
LR	KNN Wrapper	96.5684%	.953	.966
LR	LR Wrapper	96.891%	.964	.969
LR	MLP Wrapper	96.8031%	.960	.968
LR	RF Wrapper	96.4511%	.949	.965
MLP	None	98.0056%	.979	.980
MLP	DT Wrapper	96.7004%	.959	.967
MLP	KNN Wrapper	96.7004%	.957	.967
MLP	LR Wrapper	96.6711%	.956	.967
MLP	MLP Wrapper	96.979%	.964	.970
MLP	RF Wrapper	96.7591%	.959	.968
RF	None	98.8268%	.988	.988
RF	DT Wrapper	98.7975%	.987	.988
RF	KNN Wrapper	98.7975%	.988	.988
RF	LR Wrapper	98.6362%	.986	.986
RF	MLP Wrapper	98.8121%	.988	.988
RF	RF Wrapper	98.9001%	.989	.989

ORCID: Nesma Elsayed: https://orcid.org/my-orcid?orcid=0009-0004-7859-562X