Machine learning approach as an alternative tool to build a bankruptcy prediction model in banking industry

B Siswoyo\textsuperscript{1*}, N S Herman\textsuperscript{1} and D A Dewi\textsuperscript{2}

\textsuperscript{1}Faculty of Information and Communications Technology, Universiti Teknikal Malaysia Melaka (UTEM), Malaysia
\textsuperscript{2}Faculty of Information Technology, INTI International University, Nilai, Malaysia

\texttt{*bambangf1@gmail.com}

Abstract. This paper focuses on the study of a bankruptcy prediction model using a hybrid machine learning that combines two synergistic algorithms i.e. two-class boosted decision tree and multi-class decision forest. The hybrid model ensures the building of multiple decision trees whereby the latest tree corrects the previous tree, learning from the tagged data and subsequently votes on the most popular tree as the final decision of the ensemble. This hybrid machine learning is proposed to be an alternative of the bankruptcy prediction models that is able to produce three major classifications i.e. bankruptcy, grey area, and non-bankruptcy. There are five variables considered in the hybrid model which consist of working capital for total assets, retained by total asset, earnings before interest and taxes on total asset, market value of equity to total bank value of liabilities and sales of total asset. These input data are applied and tested to the public dataset produced by Bank Indonesia from year 2011-2015. The hybrid model shows a significant result whereby the overall area under curve (AUC) had successfully achieved 95% value that indicates the capability of the hybrid model to train the test data and identify the relationship of input-output data. This finding suggests that the machine learning approach can be treated as an alternative tool to build a bankruptcy prediction model for banking industry. Introduction

1. Introduction
Bankruptcy is usually interpreted as a failure of the company in carrying out the company's operations to generate profits [1]. In other words, this is related to the performance of business organizations such as banks to maintain profits. Bankruptcy is an interesting business matter to study since banking is the pulse of the economy in all countries. Nonetheless, banking in Indonesia plays a very important role, one of which is to maintain monetary stability caused by its policy on public savings and paid traffic. Concern has been expressed towards banking business performance and bankruptcies whether it constitutes a threat to economy stability [2].

There are many methods that have been applied recently to evaluate and analyze banking performance using financial statement analysis. Financial statements are company summaries that can be used as predictive knowledge [3]. Bankruptcy Prediction by Generalized Additive Models. Department of Mathematics, University of Oslo, Norway). However, financial business development has been implemented in a very complex and very dynamic environment that has led to situations where statistical assumptions about normality, homogeneity is increasingly difficult to fulfill.

Machine learning techniques have come to the picture as alternatives to simplify data analysis and find solutions [4].
Machine learning as part of artificial intelligence emphasizes pattern recognition to get optimal solutions for any case studies, including the banking industry. Machine learning for prediction has previously been used in the context of predicting bankruptcy [5]. It allows computers to find optimal data solutions automatically and accomplish relevant tasks. The idea is to undertake the tasks of machine learning algorithms, along with a comparative analysis of each model, in order to identify those that are better suited for predicting the banking performance, including bankruptcy.

Generally, machine learning works with models, including a statistical model that is designed to make prediction as accurate as possible. In this view, statistical models have limitations because of the use is too focused on building data representation and analysis to infer any relationship among variables and discover some insights. The value of statistical model is very dependent on the ability of researchers to enter the correct independent variables. In other words, if the researcher fails to identify all relevant independent variables, logistic regression has a low predictive value [6]. This statistical model had been applied to the public dataset of Bank Indonesia in comparison to the proposed hybrid model, however the results achieved less than 80% accuracy.

Another technique that has been employed by previous research is a neural network with backpropagation algorithm [7]. This algorithm used as a bankruptcy model for commercial banks in Indonesia. The results suggest that a combination of learning rate 0.7 with 2000 iterations were needed to achieve 100% accuracy within 21 seconds. The neural network model itself was able to produce 86.11% accuracy compared with the actual Bank status. This shows improvement towards the statistical model however the performance is still lesser than the proposed hybrid model.

Collectively, those results have outlined the need for an alternative tool and reliable methods for banking prediction and analysis. Hence, this paper offers new insight into the machine learning approach that combined of two algorithms to predict and evaluate banking business and bankruptcy performance. This work is inspired by research in 2017 [8]. Machine learning itself has been widely applied in various fields of business including marketing, accounting, management information systems, and production management [9]. Usually it is used to predict future stock behavior, financial crises, exchange rates, detecting credit card fraud, bankruptcy, etc.

2. Problem statements

After a series of investigations against the literature that identify the difficulties in finding an appropriate tool to build a bankruptcy model and achieve accuracy of more than 90%. This study attempts to explore whether a hybrid model (two-class boosted decision tree and multi-class decision forest) is able provide a more accurate classification of bankruptcy. In this view, the hybrid model is applied in various commercial banks in Indonesia, based on public datasets produced by Bank Indonesia in year 2011-2015. The precision and recall measurements are adopted to evaluate the performance of the proposed hybrid model in doing data classification.

3. Research objectives

The research in this paper aims to:

- Establish a bankruptcy prediction model using a hybrid model of machine learning approach and five variables as data input.
- Evaluate the hybrid model performance using precision and recall measurement.

4. Literature review

Machine learning and artificial intelligence in various fields have proven to be very useful, for example in an engineering field that has revolutionized the field of computer vision [9]. However, the role of machine learning in the economy has so far been limited.

In economics, structural models actually dominate. The reason for this is the possibility that in economics it is often more important to draw statistically significant conclusions about the causal effects of a model than to have a model with the best possible predictive performance. With that is said there is still a great potential
for machine learning in economics, such as a lot of data available and the need to understand it because many problems in economics can be considered in terms of classification or regression [5].

In addition, linear models cannot fully exploit information in data because the underlying relationships are often not linear but non-linear [10]. In this view, neural networks can predict continuous function and therefore they can be expected to provide more effective models in such applications [11].

From these advantages and disadvantages, it is clear that machine learning is suitable for problems where what is interesting does not have to be understood the governing mechanism, but rather to have a model with the best predictive performance possible. Predicting inflation is an important but difficult task faced by economists and central banks. Historically, this has been done with various types of autoregressive models. Lately, artificial neural networks, or just neural networks, have received more attention as a prediction model. There are a number of types of neural network models commonly used in the field provide pedagogical descriptions of how neural networks can be used to predict inflation [3].

Machine learning is widely used in finance to predict stock prices [12]. There have also been a number of attempts to predict currency exchange rates using machine learning methods [13]. However, in macroeconomics, the use of neural networks or even machine learning methods, in general, is not very common.

The problems of strengths and weaknesses are summarized by the bias-variance tradeoff. Bias and variance can be seen as two different sources of error. Errors due to variance are errors caused by the sensitivity of the algorithm to small fluctuations in training specified in the process of building the model (overfitting) and errors due to bias are errors caused by incorrect or inadequate assumptions of the model (under fitting) [11].

To avoid overfitting, validation sets are sometimes used in addition to training and test sets. Then prediction errors in training and validation sets can be monitored during training. However, no feedback is given to the algorithm regarding the performance of predictive models on the validation set.

Information about performance on the validation set is used to take the most common model after training [14].

Subsequently, the decision tree is a method commonly used in mining data [15]. Two-Class Boosted Decision Tree, creates a machine learning model that is based on a strengthened decision tree algorithm. The improved decision tree is an ensemble learning method in which the second tree corrects the errors of the first tree, the third tree corrects the errors of the first and second trees, and so on. Predictions are based on all tree ensembles together that make predictions.

- The decision tree driven is the easiest method for getting top performance on a variety of machine learning tasks.
- The model then selects the optimal tree using an arbitrary differentiable loss function.

On the other hand, the multi-class decision forest is known to create a machine learning model based on a forest decision algorithm. Decision forests are ensemble models that very quickly build a series of decision trees while learning from the data that is tagged. A multi-class decision forest algorithm is an ensemble learning method for classification. This algorithm works by building several decision trees and then selecting the most popular output class. Voting is a form of aggregation, where each tree is a classification decision forest produces a histogram of label frequencies that are not normalized. The aggregation process adds up this histogram and normalizes the results to get a "probability" for each label. Trees that have high predictive confidence have greater weight in the final decision of the ensemble. Decision trees are generally non-parametric models, which means they support data with a variety of distributions. In each tree, a series of simple tests are run for each class, increasing the level of the tree structure until the leaf node (decision) is reached.

Generally, the decision tree has many advantages:

- They can represent non-linear decision boundaries.
- They are efficient in computing and memory usage during training and prediction.
- They carry out the selection and classification of integrated features.
They are tough in the presence of noisy features.

5. Research methodology
Several steps are required to conduct a research methodology in this paper. Firstly, bankruptcy datasets that are generated from time-series data of five bank financial ratio variables issued at Bank Indonesia from 2011 to 2015 need to be collected and analyzed. This step called pre-processing stage. The time-series data are then converted to a size matrix [62x6] (62 rows and 6 columns). From the matrix [123x5] it has been divided into two parts, namely for training data and testing data as input into the model. Input data consists of five variables: working capital for total assets, retained by total assets, earnings before interest and taxes on total assets, market value of equity to total book value of liabilities, sales of total assets. While the output is bankruptcy, grey area, and non-bankruptcy classification.

Afterward, a filter-based feature selection is applied to segregate dataset into two partitions i.e. training dataset (70%) and test dataset (30%). Next, the development of hybrid model and tabulation of the accuracy are performed to select the highest accuracy.

After that, the five variables to calculate the bankruptcy against the real data is conducted whereby the method used for the bankruptcy classification is the proposed hybrid model. The author uses the Microsoft Azure application to perform all these tasks. The following figure 1 depicts the methodology applied in this paper.

![Figure 1. The Methodology of the proposed hybrid approach.](image)

6. Result and discussion
Based on the comparative analysis of the bankruptcy dataset, the overall results of this study demonstrate the area under the curve (AUC) for both hybrid approaches. The two-class boosted decision tree algorithm had achieved 95% of overall performance that consists of 95% accuracy, 95% precision, and 100% recall. Meanwhile, the results of the multi-class decision forest algorithm had achieved 89% of overall performance that comprises of 80% accuracy, 87% precision, 80% recall. The following figures 2 and 3 illustrate the aforementioned results whereas figure 3 describes the overall performance of the proposed hybrid method.
Figure 2. Achievement of two-class boosted decision tree.

Figure 3. Achievement of multi-class decision forest.

Figure 4. Achievement of the hybrid method.

7. Conclusion
Based on the results discussed in the previous chapter on the establishment of a bankruptcy prediction model using five input variables (working capital to total assets, retained earnings to total assets, earnings before interest and tax to total assets, market value of equity to book value of total liabilities and sales to total assets). Some conclusions can be withdrawn as the followings.

- Modeling with a machine learning approach is proven to achieve high accuracy towards the dataset of commercial banks in Indonesia.
- The five variables can model used as input parameters and the behavior between variable are reasonable to be analyzed as finance ratio to bankruptcy performance.

References
[1] Brahmana R K 2007 Identifying financial distress condition in Indonesia manufacture industry United Kingdom: Birmingham Business School, University of Birmingham,
[2] Altman E I 1968 Financial ratios, discriminant analysis and the prediction of corporate bankruptcy The Journal of Finance 23(4) 589–609
[3] Berg D 2005 Bankruptcy Prediction by Generalized Additive Models Statistical Research Report
1, January 2005 Norway: Department of Mathematics, University of Oslo

[4] G Eason et al. 2011 Predicting Bankruptcy with Robust Logistic Regression Journal of Data Science 9 565-584

[5] Pedregosa, et al. 2011 Scikit-Learn: Machine Learning in Python Journal of Machine Learning Research 12 2825-2830

[6] Aryati and Manao 2000 Financial Ratios as Predictors of Problem Banks in Indonesia National Accounting Seminar Jakarta

[7] Malaka and Hartojo 2012 Bankruptcy Prediction Models Commercial banks in Indonesia use the Back Propagation algorithm

[8] Matson B and Steinert O Department Of Economic University of Gothenburg School of Business Economics and Law Gothenburg Sweden

[9] Inga S and Bronius N 2008 Conceptual Model of Intelligent University Economic Engineering 4 59

[10] Zhou L, Lai K K and Yen J 2014 Bankruptcy prediction using SVM models with a new approach to combine features selection and parameter optimisation International Journal of Systems Science 45(3) 241–253

[11] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15(1) 1929-1958

[12] Hung C et al. 2006 Hybrid Probability Based Ensembles For Bankruptcy Prediction Hybrid Probability Based Ensembles For Bankruptcy Prediction. Department of Management Information Systems, Chung Yuan Christian University Taiwan ROC, School of Computing and Technology, University of Sunderland, UK

[13] Doganay M M, Ceylan N B and Aktas R 2006 Predicting financial failure of the Turkish banks Annals of Financial Economics 197–117

[14] Giang H 2005 Using Artificial Neural Networks in Predicting Company Failures Journal. Social Sciences

[15] Rokach L and Maimon O 2008 Data mining with decision trees: theory and application World Scientific Pub Co Inc.