Firm Performance Prediction for Macroeconomic Diffusion Index using Machine Learning

Cu Nguyen Giap[1], Dinh Thi Ha[3]
Vu Quang Huy[4], Do Thi Thu Hien[5]
Faculty of Economic Information System and e-Commerce
Thuongmai University
Hanoi, Vietnam

Dao The Son[2], Le Mai Trang[6]
Department of Economics
Thuongmai University
Hanoi, Vietnam

***Abstract*—Utilizing firm performance in the prediction of macroeconomic conditions is an interesting research trend with increasing momentum that supports to build nowcasting and early warning systems for macroeconomic management. Firm-level data is normally high volume, with which the traditional statistics-based prediction models are inefficient. This study, therefore, attempts to assess achievements of Machine Learning on firm performance prediction and proposes an emerging idea of applying it for macroeconomic prediction. Inspired by "micro-meso-macro" framework, this study compares different machine learning algorithms on each Vietnamese firm group categorized by the Vietnamese Industry Classification Standard. This approach figures out the most suitable classifier for each group that has specific characteristics itself. Then, selected classifiers are used to predict firms' performance in the short term, where data was collected in wide range enterprise surveys conducted by the General Statistics Office of Vietnam. Experiments showed that Random Forest and J48 outperfomed other ML algorithms. The prediction result presents the fluctuation of firms' performance across industries, and it supports to build a diffusion index that is a potential early warning indicator for macroeconomic management.**

Keywords*—Firm performance prediction; machine learning algorithms; diffusion index

I. Introduction

Macroeconomic situation is always an important factor for all economic sectors, and it is trivial that macroeconomic forecasting is very important [1]. From the necessity of macroeconomic forecasting, there have been many studies on this issue, such as predicting GDP [2], forecasting inflation [3], unemployment [4], exchange rates [5]. These indicators are also predicted in different aspects, such as forecasting growth level [6], or degree of fluctuation [7].

Faced with new practical demands, due to the re-emergence of economic crises, economists are more interested in alerting abnormal situations instead of just giving out the predicted value of indicators. And along with the improvement in the availability of input data of prediction problem, new achievements in data processing methods, and computing power, a new group of studies with the concept of "nowcasting" and early warning system in macroeconomic arises. Aastveit, Gerdrup et al. used big data and machine learning techniques to forecast real-time GDP [8]. The large data was used to report and forecast macroeconomic situation [9], and Galbraith and Tkacz applied the GDP reporting method with electronic payment data [10]. On the other side, Reinhart et al. [11] studied the development of early warning models of financial market risk assessment for emerging markets. Ciarlone and Trebeschi (2005) studied how to design an early warning system for debt crises [12], and Sevimet al. (2014) built an early warning system to forecast currency crises [13].

In most cases, the input data used for macroeconomic prediction are aggregate economic indicators, such as consumer price index (CPI) or gross domestic product (GDP). However, firm performance is an indispensable and crucial factor that needs to be considered when predicting macroeconomic conditions. In recent decades, there is a new trend of using firm-level data in predicting macroeconomic aspects [14]–[20]. The researchers have proved the importance of using firm-level information in macroeconomic prediction from both theoretical and practical points. From the theoretical point, a framework to research the effect of micro-foundations information in macro-economic aspect was proposed in [19], [20]. In practical, Joao et al have using firm performance to conduct new micro-aggregated factor that was proved to be useful in GDP prediction [14]. Productivity in firm-level data was also used to build a micro-aggregated factor that is similar to total factor productivity (TFP) and this micro-aggregated factor is facilitated in the prediction of other macro-economic indicator models.

Using firm-level data on macroeconomic prediction brings several advantages. a) micro-aggregated series presents the dynamics of the published aggregate factors reasonably. b) micro-foundation information can identify the factors underlying the differences in the efficiency of all manufacturing. c) firm-level information is measured in high frequency based on information communication technology [14]–[16], [18]. This approach also brings a great opportunity for improving qualities of prediction models thank to data's granularity. However, it is also posing a challenge because firm-level data is normally high-volume data. With high volume data, traditional statistic-based prediction models are inefficient. In proposals of Bartelsman et al, and Brito et al, had to build a microaggregate factors before using them in prediction models. In opposite, machine learning algorithms become the highly potential methods for macroeconomic state prediction based on firm-level information directly.

However, it can be concluded that until now, there has been no study which systematically evaluates the performance

of ML algorithms on this issue. From this point of view, this article uses ML algorithms to predict firms' performance based on firm survey data and other additional information. This study differs from other studies in the way that firm performance prediction is used to support for macroeconomic perspective, rather than for microeconomic management. For this purpose, a huge data including information of a wide range of companies must be processed, and the firms' information used is public information only. Using secret firms' information for macroeconomic analysis is inappropriate. Besides, the companies of each industry group have exclusive/specific characteristics, therefore the suitable models of each group would not be the same. Inspired by "Micro-meso-macro" framework [19], therefore, this study aims to build the different predictive classifier for the firm's performance of different industry groups. This article also aims to build the warning of the macroeconomic situation based on firms' performance classification.

Firm's performance is measured by different indicators including financial indicators and market signal indicators. This article uses the two most popular indicators for the firm's performance, which are the return on asset (ROA) and the return on equity (ROE). Predicting other firm performance indicators [21] is going to be a research question for the next study.

Macroeconomic forecasting in Vietnam has been implemented by the government for a long time. However, it was not until 1984 that Vietnam had the GDP index for the first time, which is the most basic indicator of the macroeconomy. There are some forecasts of basic macroeconomic indicators such as GDP and inflation. However, there is no research applies the same approach with this study for Vietnamese macroeconomic forecasting.

In this article, the information of Vietnamese enterprises collected by the General Statistics Office of Vietnam in the 2010-2015 period has experimented. The result shows the potential application of selected machine learning algorithm to supply important information: predictive firms' performance for the overall economy, which can indicate macroeconomic condition. A proposal micro-aggregate factor, akin to diffusion index, built from firm performance classification prediction and its potential use are shown in this research.

In the remaining part of this paper, section 2 presents the related works. In section 3, the research methodology and preliminary are introduced. Section 4 presents the experiments and evaluation of the performance of algorithms on Vietnamese firm's data sets. Finally, Section 5 presents some conclusions, discussion, and research extension in the near future.

II. Related Works

Utilization of Firm Performance Prediction (FPP) for macroeconomic perspective based on Machine learning algorithms is still an ongoing research area. There are few publishes proposed the clear application of FPP for policymakers. However, machine learning algorithms have been widely applied for firms' performance prediction for other purposes [22]–[32], and these models will be also useful for the government's approaches on macroeconomic management.

Different novel models for FPP were proposed on each application such as bankruptcy, financial rating, financial distress, business failure, and so on. These applications consider different aspects of FPP, therefore its use different FPs indicators and independent variables. In the beginning, the researchers and managers have strongly focused on FFP methods using the accounting factors, then market-based indicators have been concerned also. Recently, new technological approaches have used sentiment-based analysis to improve the firms' performance prediction [33]–[35].

In firm performance prediction, successive machine learning algorithms including decision tree, support vector machine, artificial neural network, and its modification are well-known, and the remarkable studies are briefly summarized herein.

Decision tree included boosting the algorithm, the random forest is a well-known machine learning algorithm for FPP [36]–[41]. Delen deeply researched on applying decision tree for firm performance evaluation [36]. He studied common decision tree models including CHAID, C5.0, QUEST, and C&RT for the large and rich feature dataset and proposed an application framework for FPP using a decision tree with financial factors. Bankruptcy prediction is also well studied using decision tree model. Zibanezhad and Foroghi [37] extracted 25 Financial ratios and used these ratios as independent variables for building a Classification And Regression Tree. This work adapted very well to continuous financial ratios data and showed the applicable of C&R tree for bankruptcy prediction problem. Recently, Jardin [38] proposed a novel application of decision tree for firm financial evolution prediction in mid-term, and bankruptcy. His work made an improved approach by using closed time-series data of six financial dimensions to make his model become sensitive to short-term changes. Many other works used decision tree for firm performance issues were introduced, Bastı [40] and Yeo [41] were good examples.

Support vector machine (SVM) is a supervised learning algorithm developed by Vapnik [42] and is successfully applied for classification problems. In the realm of finance applications, SVM has been applied in a various problems such as selecting bankruptcy predictors [43], stock price index predicting by financial time series data [44]–[46], forecasting bankruptcy or failure abilities [31], [47]–[50]. In those aspects, SVM approaches were compared with many other methods and this algorithm showed its potential ability. Francis [46] compared SVM with back-propagation neural network (BPNN) based on criteria such as normalized mean square error (NMSE), mean absolute error (MAE), directional symmetry (DS) and weighted directional symmetry (WDS). SVM is also assessed about the predictive performance along with some traditional approaches including the Linear Discriminant Classifier (LDA), Multi-layer Perceptron (MLP), and the Learning Vector Quantization (LVQ) [43]. Min [47] evaluated the SVMs with MDA, logistics, and BPNN or Wu [48] compared a genetic algorithm based SVMs with DA, logit, probit, and ANNs. In these research, the results

indicated that SVM performed better than other approaches with higher percentage classified correctly, or higher precision and lower error rates. However, there are some studies recently proved that ANN models to be superior to SVM approaches with smaller estimated relative error costs [50].

Many proposal ANN models has been applied to finance problems especially in bankruptcy and financial distress [24], [29], [32], [51], since 1990's. Odom and Sharda were pioneers in applying ANNs to predict the failures of firms [52]. They developed an ANN model for bankruptcy prediction and compared proposal model to multivariate discriminant analysis (MDA) by empirical tests using financial data from various companies. Tam [53] applied backpropagation neural network (BPNN) for bankruptcy prediction of banks and compared with existing models include discriminant analysis (DA), factor-logistic, K- nearest neighbor (k-NN) and decision tree ID3. The results showed that ANNs offered better predictive accuracy than other models. Coats and Fant [54] proposed an ANN model as an alternative analysis method of the same ratios used by MDA. They showed the ANN approach was more effective than MDA. Bell and T.B [55] compared logistic regression and ANNs in predicting bank failures. Their results indicated that both methods having similar predictive accuracy. Besides, there were many research combining ANN with other soft computing techniques such as fuzzy sets [56], genetic algorithm [57], rough sets [58].

To build a framework that uses machine learning algorithms to predict Vietnamese FP for macroeconomic perspective, those algorithms are tested and compared with different common machine algorithms to identify the best model for each industry group.

III. Research Methodology

A. Classification Model Construction

This study aims to predict performance at firm-level data with the purpose of supporting forecasting at macroeconomic level. For this, a huge data including information of enterprises across the economy is processed. For privacy and practical isssues, only publictly available information of firms' performance is used. Following the framework introduced by Dopfer [19], this study aims to build the different predictive classifier for the firm's performance of different industry groups following Vietnam Standard Industrial Classification. The results of classifiers are used to estimate a proposed index that is a trigger in an early warning system for macroeconomic management.

The study uses a research methodology which includes two critical stages, the first stage collects and preprocesses the large raw data set, and the second stage builds the typical model for each industry group. The specific steps of this research methodology are depicted in Figure 1.

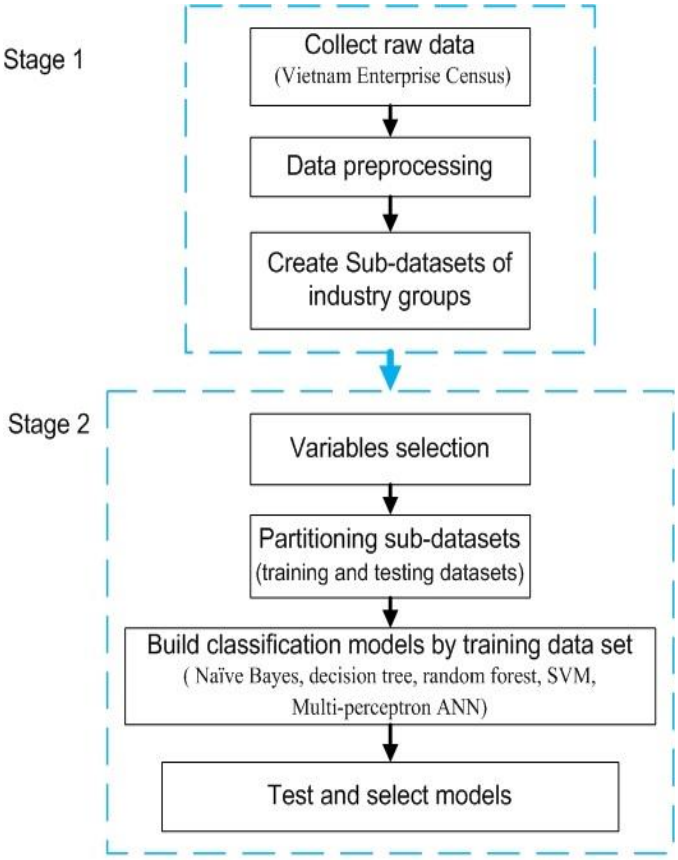

Fig. 1. Steps followed in Research Methodology.

The first stage has three steps: (1) raw data collection, (2) preprocessing, and (3) sub-setting raw data to each industry group. The output of this stage is valid datasets for stage 2 that builds suitable prediction models. In the first step, the raw dataset of firms' information is collected from GSO's annual economic census. In the second step, outliers, are removed. Missing values are either removed or replaced by the average value of variables it belongs to. Data noises are processed to provide an acceptable quality dataset to later stages. In the third step, the data set is partitioned by industry group code according to Vietnam Standard Industrial Classification, a slightly modified version of the Global Industry Classification Standard, in the third step. And then, the data sets belong to industry groups are validated before being passed on to the second stage.

In the second stage, each sub-dataset is split into training and testing datasets by a specific proportion in the first step. Then, the training dataset is used to train different ML approaches including Naïve Bayes, decision tree (J48), random forest, SVM, ANN. These models are tested on the testing dataset and the models' performances are evaluated by different measurements containing precision, recall, accuracy, ROC. The testing result is analyzed to make the insight of using ML approaches for FPP in macroeconomic perspective.

1) Naïve bayes classifier: Naïve Bayes is a common machine learning technique that is developed based on the Bayes's theorem and it is suited when the dimensionality of the inputs is high and assume that inputs are independent is satisfied. Naïve Bayes classifier takes input instance as a feature vector x={x1, …, xn} and classes dependent variable y by posterior probability Prob(yi|x) where yi is a possible outcome of y. Naïve Bayes classifier is commonly trained by a supervised method such as Maximum-likelihood on a given training set [59]. Particularly:

Posterior probability Prob(y_i|x)

$$prob(y_i \mid x) = \frac{prob(y_i)\, prob(x_1,...,x_n \vee y_i)}{prob(x_1,...,x_n)}$$

In practice, $prob(y_i)$ and $prob(x_1,...,x_n)$ are calculated from a training set directly and $prob(x_1,...,x_n \mid y_i)$ is equivalent to the joint probability model $prob(x_1,...,x_n,y_i)$. Applying the chain rule and Naive independent assumption of inputs, there is:

$$prob(x_1,...,x_n,y_i) = prob(x_1 \mid x_2,...,x_n,y_i)\, prob(x_2,...,x_n,y_i)$$
$$= prob(x_1 \mid x_2,...,x_n,y_i)\, prob(x_2 \mid x_3,...,x_n,y_i)\, prob(x_3,...,x_n,y_i)$$
$$= prob(x_1 \mid x_2,...,x_n,y_i)\, prob(x_2 \mid x_3,...,x_n,y_i)...$$
$$prob(x_{n-1},x_n,y_i)\, prob(x_n,y_i)\, prob(y_i)$$
$$= prob(x_1 \mid y_i)\, prob(x_2 \mid y_i)...prob(x_{n-1} \mid y_i)\, prob(x_n \mid y_i)\, prob(y_i)$$
$$= prob(y_i)\prod_{j=1}^{n} prob(x_j \mid y_i)$$

Because $prob(x_j \mid y_i)$ is proportional to $prob(x_1,...,x_n,y_i)$, henceforth, the classifier model of dependent variable y to a specific class $\hat{y}$ is:

$$\hat{y} = \underset{i \in \{1,...,k\}}{\arg\max}\ prob(y_i) \prod_{j=1}^{n} prob(x_j \mid y_i)$$

2) Decision tree: Decision tree (DT) is a tree-based structure model that expresses the possible consequent states with its chance events and outputs. Decision tree is a non-parametric supervised machine learning algorithm that is used popularly for classification and regression problems. Learning a decision tree is based on partitioning the set of training examples into smaller and smaller subsets where each subset is as "pure" as possible. The purity for a particular subset is measured according to the number of training samples in that subset having the same class label. In practice, a DT structure is constructed directly from a training set by an iterative process that starts with a null root node and repeatedly split a node on an attribute-based on information gain. A DT is built by C4.5, J48, C5.0 algorithms, and they all follow the same recursive process extends from Quinlan's earlier ID3 algorithm.

The most important task in decision tree construction process is finding the normalized information gain from splitting on an attribute. This is complete by following steps [60].

Calculate information conveyed by a distribution of the set of classes P in a current dataset, also called the Entropy of P, is:

$$I(p) = -(p_1 \log(p_1) + p_2 \log(p_2) + ... + p_n \log(p_n))$$

Calculate the incremental information by partition a set T on a non-categorical attribute X into sets T_1, T_2, .., T_m

$$Info(X,T) = \sum_{i=1}^{m} \frac{|T_i|}{|T|} * I(T_i)$$

Calculate the quantity Gain(X,T) defined as is the measure of the difference in entropy from before to after the set T is split on an attribute X{\displaystyle A}:

$$Gain(X,T) = I(T) - Info(X,T)$$

3) Random forest: Random forest is a resampling approach for classification and regression problems. Random forest builds a classifier by assembling individual simple classifiers trained on different sub-datasets generated by bootstrapping a training set [41]. Random forest classifier improves the quality of a classifier built on a single decision tree by solving overfit and bias problems. Random forest uses Bagging (bootstrap aggregating) algorithm, which uses multiple versions of the training set, each created by bootstrapping the original training data to train the models. Each of these bootstrap data sets is used to train different component classifiers that are simple decision trees commonly, and then a final classification decision is form by a voting process of each component classifier.

4) Support vector machine: Recent years, support vector machine (SVM) is a new succeeded supervised learning algorithm for classification problems including the firm's performance prediction problem [47]. SVM is a non-parametric algorithm aimed to find the optimal hyperplanes that separate classes on a training dataset. An SVM is trained by solving a large quadratic programming problem. For the proposed problem, SVM is trained by a sequential minimal optimization algorithm, in which a complex quadratic programming problem is broken into a series of smallest possible quadratic programming problems and these small quadratic programming problems are solved analytically.

For a binary classification problem with a dataset $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, where x_i is an input vector and y_i is a relative output label. A soft-margin support vector machine is trained by solving a quadratic programming problem:

$$\max_a \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j K(x_i,x_j) a_i a_j$$

Subject to:

$$0 \le a_i \le C, i = 1, 2, ..., n$$

$$\sum_{i=1}^{n} y_i a_i = 0$$

Where C is a regularization of SVM model and $K(x, y)$ is a kernel function that can be linear, polynomial, or exponential kernel.

5) Feed-Forward Artificial Neural Network (ANN): Feed-forward ANN is a great machine learning model that can represent a highly complex relationship between inputs and outputs. Feed-forward ANN has three layers of architecture includes input layer, hidden layers, an output layer, and the connections of nodes are represented by an acyclic directed graph. One node of ANN has multiple inputs, a weight vector *w* and one output with a bias, as depicted in Fig.2 [54].

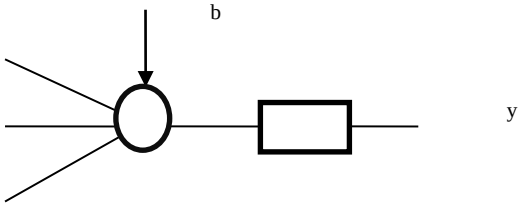

Fig. 2. A Neuron of Dynamic Structural ANN.

In this study, the ANN uses sigmoid activation function is defined below:

$$\varphi(net) = \frac{1}{1 + e^{-net}}$$

In which, for the node j

$$net_j = b + \sum_{i=1}^{n} x_i w_{ij}$$

In the input layer, each input node is linked with only one input and this node has connections to all hidden nodes. The hidden layer is a set of consequent layers or is one full connect layer only. In any case, a hidden node connects to all nodes in the next layer, and in the last hidden layer, a hidden node has direct connections to all nodes of the output layer. The number of nodes in the output layer is the same with the number of outputs.

Feed-forward ANN is commonly trained by Back – Propagation algorithm, in which the weights of nodes are changed with subject to minimize output error.

$$\varepsilon = \sum_{i=1}^{n} (t_i - y_i)^2$$

where: t_i is an instance output value in training set and y_i is its relative network output**.**

B. A proposal on a Diffusion Index as an early Warning System of Real Business Cycle

At this stage, our prediction model is good at forecasting firms' ROA classification. Using input data from the beginning of each year, firms are predicted how well they will perform at the end of that year and classified into each of the five categories from lowest ROA to highest ROA. Each firm then can be predicted to stay in the same class, move up, or move down. We propose to use a diffusion index-based computed indicator to describe the overall situation of firms across industries in the economy by looking at the number of firms that are performing better or worse. Diffusion indexes was formally mentioned in [61] and it has been applied regularly since then, such as for macroeconomic forecasting [62].

$$DI_{fp} = \frac{1 * N_{mh} + 0.5 * N_{ss} + 0 * N_{ml}}{N_{mh} + N_{ss} + N_{ml}} * 100$$

Where

- DI_{fp} is the diffusion index based on firm performance at an time point.
- N_{mh} is the number of firm moving to higher classes: note that we compare the end-of-year predicted value to the begin-of-year actual value.
- N_{ss} is the number of firm that stay in the same class.
- N_{ml} is the number of firm moving to lower classes.

The index can vary from 0 to 100. If the index is equal to 50, it can be consider as averagely all firms are performing in the same classes. If the index is higher than 50, more firms moving up them firms moving down and that's a signal of an improving macroeconomic performance. If the index is higher lower than 50, more firms moving down than firms moving up and that's a signal of a slowing down macroeconomic performance.

This framework is similar to the use of the Purchasing Manager Index that has been widely used as a leading indicators to forecast the economy [63]. However, the index proposed in this study has a significant advange of computing from actual firms' performance across the economy rather than from getting opinions from purchasing managers of selected enterprises.

Although the data used in this expriment study was the enterprise census implemented annually, the main purpose of showing the potential use of large data on firms' information is still valid. In practice and further investigation, the government with its ability to assess firms' information at higher frequency or sometime even real time, can apply the same framework. And given that this study tries to predict firms' performance in subset of industries will also allow the government to create diffusion indexes for each industry or sector, therefore having able to detect the macroeconomic performance with more details.

IV. Experimental Result

A. Data Description and Variables Selection

This study applies and compares different well-known machine learning algorithms on the data of Vietnamese enterprises collected by the General Statistics Office of Vietnam in the 2010-2015 period. This dataset was carried out with all firms in every type of business and every industry in Vietnam. It includes more than 500,000 companies with nearly 200 variables for each observation. These variables reflect different aspects of firms such as business tax ID, type of business, asset and capital structure, employee structure, the result of business, and others.

These are annual enterprise census conducted by the General Statistic Office of Vietnam to provide the government with information about the firms' performances. These surveys have some disadvantages for the application of machine learning algorithms because the questionnaire is designed with closed-ended questions for common statistical analysis. Despite several limitations to the performance of machine learning algorithms, it is still the best existing dataset with information of firm performance for macroeconomic perspective, and therefore suitable for this research.

As mentioned in the literature review, common theory of firm performance evaluation used many indicators, but in this emerging study, two indicators including ROE and ROA are used to fulfill the study targets. Variables in the raw dataset from the General Statistics Office of Vietnam which do not support ROE, ROA prediction are removed. On the other hand, some new variables are generated from original variables to use in prediction models. The final selected variables are described in Table 1, in which ROA, ROE are output indicators whereas other variables are input indicators. All these input parameters are values at the begin of each year and they are used for predicting firm performance indicators at the end the of year.

In addition, this study aims to assess not only the different impacts of indicators on the prediction model but also the difference between industry groups, therefore the dataset is divided into subsets corresponding with industries. This division is performed according to codes of each industry in Decision on Vietnam Standard Industrial Classification of Prime Minister, No. 27/2018/QD-TTg. Although there are 99 industries in Vietnam in this division, the study select test on industries that have the number of firms larger than 20,000 to guarantee learning quality of prediction models. These selected industries are described in Table 2.

TABLE I. Indicators of Research Model

Index	Indicator	Index	Indicator
1	Type of business	16	Year-opening debt ratio [total debt/total capital]
2	Business size [1=SME, 2=big]	17	Export value
3	State-ownership status	18	Gross revenue (VND million)
4	Total assets (VND million)	19	Net revenue
5	Total equity (VND million)	20	Core-business Gross revenue
6	Total debt (VND million)	21	Ratio of core business revenue to gross revenue (%)
7	Total number of employees	22	Percentage of core-business Gross revenue
8	Total number of female employees	23	Profit before tax
9	Total number of core-business employees	24	Profit after tax
10	Number of change employees	25	Business tax
11	Percentage of female employees (%)	26	Total of earning (VND million)
12	Percentage of core-business employees (%)	27	Year performing the survey
13	Second-industries status	28	Return on Asset
14	Percentage of state shares (%)	29	Return on Equity
15	Owner's equity ratio (%)		

TABLE II. THE CODES OF SELECTED INDUSTRIES

No	Code	Description
1	01	Agriculture and related service activities
2	10	Manufacture of food products
3	14	Manufacture of wearing apparel
4	42	Construction of civil engineering structures
5	43	Specialized construction activities
6	47	Retail trade, (except motor vehicles, motorcycles
7	49	Land transport and transport via railways and via pipelines
8	55	Accommodation
9	56	Food and beverage service activities
10	68	Real estate activities
11	73	Advertising and market research

B. Experiment Result

Return on asset (ROA) and return on equity are two numerous variables in raw survey data set. However, the light changes of ROA, ROE are meaningful for managers of firms but not for the policymakers. Therefore, this study transforms ROA, ROE variables into two categorical variables by the specific interval border. The ROA is discretized by intervals $\{(,0],(0,5],(5,20],(20,30],(30,)\}$ and ROE is discretized by intervals $\{(,2.5],(2.5,7.5],(7.5,15],(15,20],(20,)\}$. These segments are encoded by class labels as 0, 1, 2, 3 and 4. Table 3 and Table 4 show the number of firms belongs to different classes of every industry group.

In general, the performances of five well-known machine learning algorithms, Naïve Bayes, decision tree (j48), random forest, SVM, MLP, for Vietnamese firms' performance prediction is shown in figure 3 and 4.

According to the proportion of correctly classified instances, as shown in figures 3 and 4, in both cases, two tree-based algorithms including J48 and random-forest have out-performed other algorithms. In the case of ROA prediction, J48 algorithm has the minimum proportion of correctly classified instances is 86.31% for No.56 industry and reach maximum proportion at 95.77% for No.42 industry, and the average proportion is 91.61%. The random forest algorithm is even better, it has the minimum proportion of correctly classified instances is 87.38% for No.56 industry and reach maximum proportion at 95.77% for No.42 industry, and the average proportion is 91.81%. Two algorithm SVM and MLP have close performance and they only work well in several industry groups including No.42, No.43, and No.49 with the proportions of correctly classified instances are higher than 80%. Naïve Bayes algorithm doesn't work in this case. It has proportions of correctly classified instances from 6.92% to 44.21%.

TABLE III. NUMBER OF FIRMS IN EACH CLASS AND INDUSTRY GROUP BY ROA CLASSIFICATION

Class\industry code	1	10	14	42	43	47	49	55	56	68	73
0	701	289	62	784	1,248	3,946	607	41	9	405	375
1	3,912	2,565	2,090	7,955	3,959	12,823	4,634	2,567	1,737	1,850	2,211
2	1,677	767	693	677	465	3,135	689	631	634	375	514
3	237	116	109	71	53	388	60	72	100	78	95
4	473	163	146	113	75	608	110	289	120	92	105
Total	**7000**	**3900**	**3100**	**9600**	**5800**	**20900**	**6100**	**3600**	**2600**	**2800**	**3300**

TABLE IV. NUMBER OF FIRMS IN EACH CLASS AND INDUSTRY GROUP BY ROE CLASSIFICATION

Class\Industry code	**1**	**10**	**14**	**42**	**43**	**47**	**49**	**55**	**56**	**68**	**73**
0	3,413	1,680	1,194	6,476	4,038	11,785	3,630	2,075	1,073	1,759	1,605
1	1,839	908	693	1,727	954	4,465	1,339	847	701	406	754
2	836	515	440	713	361	2,174	543	367	405	265	369
3	265	189	181	219	105	717	174	107	108	90	121
4	647	608	592	465	342	1,759	414	204	313	280	451
Total	**7000**	**3900**	**3100**	**9600**	**5800**	**20900**	**6100**	**3600**	**2600**	**2800**	**3300**

- Result of ROA prediction:

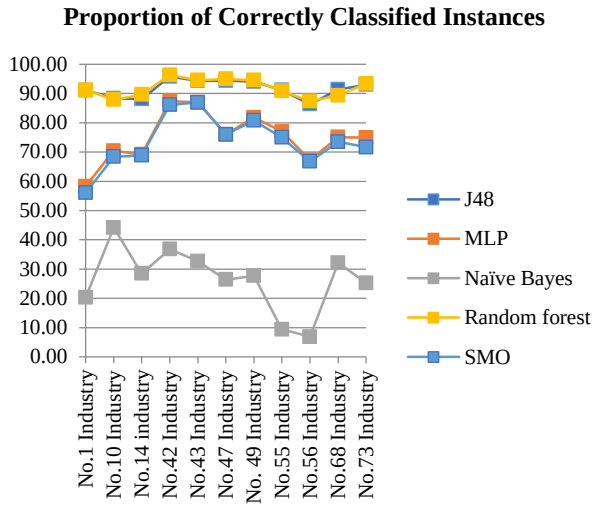

Fig. 3. The Proportion of Correctly Classified Instances of ROA.

In case of ROE prediction, J48 algorithm takes the lead with a minimum proportion of correctly classified instances is 86.31% for No.56 industry and reaches maximum proportion at 95.77% for No.42 industry, and the average proportion is 91.48%. The random forest algorithm follows with a minimum proportion of correctly classified instances is 80.05 % for No.10 industry and reaches maximum proportion at 91.48% for No.42 industry, and the average proportion is 86.37%. Three remain algorithms, Naïve Bayes, SVM, and MLP have the close performance in this case. SVM algorithm is a little better with average proportions of correctly classified instances is 55.82% while MLP reaches 54.00% and Naïve Bayes has 49.51%.

- Result of ROE prediction:

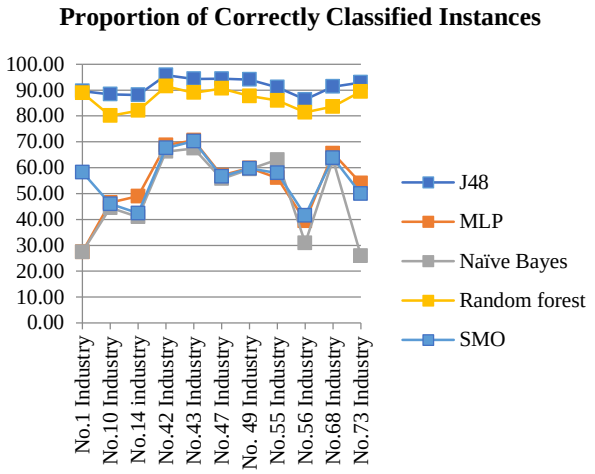

Fig. 4. The Proportion of Correctly Classified Instances of ROE.

1) Performance of ML algorithms for all industry groups: To evaluate the performance of every selected algorithm in all industry, three measures are used including precision, recall, and ROC-area. The tables below show the performances of all algorithms, and the bold numbers show the best results of each measurement on each industry.

The tables 5-8 show that in both cases, regarding the ROC-area the random forest algorithm outperforms all other algorithms and J48 algorithm follows closely. In the case of ROA prediction, the random forest algorithm is better than J48 according to recall and precision (Table 5). However, J48 is better than random forest according to recall and precision in ROE prediction (Table 7). In this study, the random forest algorithm is the best for ROA prediction problem and J48 is chosen for ROE prediction problem.

- Result of ROA prediction:

TABLE V. THE PERFORMANCES OF NAÏVE BAYES, DECISION TREE AND RANDOM FOREST

Index	Code/ Division	Algorithms								
		Naïve Bayes			*Decision tree*			*Random Forest*		
		Precision	*Recall*	*ROC-area*	*Precision*	*Recall*	*ROC-area*	*Precision*	*Recall*	*ROC-area*
1	01	0.436	0.203	0.582	**0.911**	0.911	0.953	0.906	**0.912**	**0.988**
2	10	0.625	0.442	0.762	**0.879**	0.883	0.929	0.865	**0.879**	**0.976**
3	14	0.517	0.285	0.512	0.879	0.881	0.913	**0.898**	**0.896**	**0.979**
4	42	0.767	0.368	0.673	0.956	0.958	0.951	**0.960**	**0.964**	**0.994**
5	43	0.684	0.327	0.797	**0.941**	0.943	0.964	0.940	**0.946**	**0.994**
6	47	0.534	0.264	0.701	0.944	0.944	0.974	**0.948**	**0.950**	**0.995**
7	49	0.632	0.277	0.634	**0.940**	0.940	0.959	0.938	**0.945**	**0.993**
8	55	0.592	0.094	0.576	**0.913**	**0.911**	0.944	0.900	0.908	**0.982**
9	56	0.523	0.069	0.529	0.861	0.863	0.913	**0.865**	**0.874**	**0.971**
10	68	0.524	0.321	0.653	**0.912**	**0.914**	0.951	0.879	0.893	**0.983**
11	73	0.416	0.252	0.584	0.928	0.930	0.965	**0.934**	**0.934**	**0.992**

TABLE VI. THE PERFORMANCES OF SVM AND MLP

Index	Code/ Division	Algorithms					
		Support Vector Machine			*Multi-perceptron*		
		Precision	*Recall*	*ROC-area*	*Precision*	*Recall*	*ROC-area*
1	01	?	0.560	0.553	?	0.583	0.668
2	10	?	0.684	0.616	?	0.704	0.74
3	14	1	0.021	0.575	0.212	0.075	0.765
4	42	?	0.861	0.710	?	0.874	0.808
5	43	?	0.869	0.873	?	0.869	0.873
6	47	?	0.760	0.815	?	0.760	0.815
7	49	?	0.808	0.703	?	0.818	0.765
8	55	?	0.749	0.637	?	0.770	0.731
9	56	?	0.667	0.506	?	0.672	0.636
10	68	?	0.735	0.691	?	0.751	0.807
11	73	?	0.716	0.652	?	0.749	0.817

- Result of ROE prediction:

TABLE VII. THE PERFORMANCE OF NAÏVE BAYES, DECISION TREE AND RANDOM FOREST

Index	Code/ Division	Algorithms								
		Naïve Bayes			*Decision tree*			*Random Forest*		
		Precision	*Recall*	*ROC-area*	*Precision*	*Recall*	*ROC-area*	*Precision*	*Recall*	*ROC-area*
1	01	0.336	0.274	0.557	**0.892**	**0.893**	0.954	0.884	0.889	**0.986**
2	10	0.357	0.445	0.648	**0.834**	**0.833**	0.936	0.786	0.801	**0.960**
3	14	0.341	0.410	0.633	**0.848**	**0.851**	0.934	0.812	0.820	**0.968**
4	42	0.561	0.662	0.672	0.913	0.913	0.956	**0.914**	**0.915**	**0.989**
5	43	0.586	0.674	0.742	**0.908**	**0.907**	0.962	0.881	0.890	**0.984**
6	47	0.432	0.557	0.643	**0.914**	**0.914**	0.964	0.904	0.906	**0.989**
7	49	0.499	0.594	0.633	**0.897**	**0.895**	0.955	0.872	0.877	**0.983**
8	55	0.666	0.631	0.858	**0.880**	**0.882**	0.942	0.853	0.859	**0.977**
9	56	0.433	0.309	0.565	**0.838**	**0.837**	0.917	0.804	0.814	**0.961**
10	68	0.542	0.631	0.765	**0.889**	**0.889**	0.954	0.821	0.836	**0.975**
11	73	0.525	0.259	0.698	**0.896**	**0.896**	0.949	0.894	0.894	**0.987**

TABLE VIII. THE PERFORMANCES OF SVM AND MLP

Index	Code/ Division	Algorithms					
		Support Vector Machine			*Multi-perceptron*		
		Precision	***Recall***	***ROC-area***	***Precision***	***Recall***	***ROC-area***
1	01	?	0.583	0.668	?	0.478	0.587
2	10	?	0.460	0.572	?	0.465	0.656
3	14	0.462	0.267	0.615	0.478	0.490	0.736
4	42	?	0.676	0.524	?	0.688	0.708
5	43	?	0.702	0.534	?	0.706	0.742
6	47	0.449	0.566	0.508	?	0.570	0.686
7	49	?	0.598	0.508	?	0.598	0.628
8	55	?	0.579	0.510	?	0.561	0.611
9	56	?	0.415	0.508	?	0.394	0.603
10	68	?	0.638	0.561	?	0.655	0.773
11	73	?	0.499	0.522	?	0.540	0.710

2) Performance of best ML algorithms for each class: Deeply analyze the efficiency of the best algorithms in both ROA and ROE prediction problems, the algorithms' performances on different classes are depicted in the following tables and figures.

- Result of ROA prediction is presented in Table 9 and Figure 5:

TABLE IX. THE PERFORMANCE OF RANDOM FOREST ALGORITHM

Class label	0	1	2	3	4
No.1 Industry	99.86	96.86	87.18	19.83	81.82
No.10 Industry	99.31	97.58	73.92	8.62	36.81
No.14 industry	96.77	96.79	84.70	27.52	53.42
No.42 Industry	99.36	99.64	70.31	16.90	50.44
No.43 Industry	100.00	98.56	65.59	7.55	33.33
No.47 Industry	99.80	97.99	86.86	42.27	75.16
No. 49 Industry	99.84	98.99	76.78	6.67	36.36
No.55 Industry	68.29	97.51	78.76	9.72	80.97
No.56 Industry	55.56	95.51	82.18	16.00	59.17
No.68 Industry	99.75	98.11	64.80	8.97	32.61
No.73 Industry	100.00	98.10	84.82	29.47	70.48
Average	**92.59**	**97.78**	**77.81**	**17.59**	**55.51**

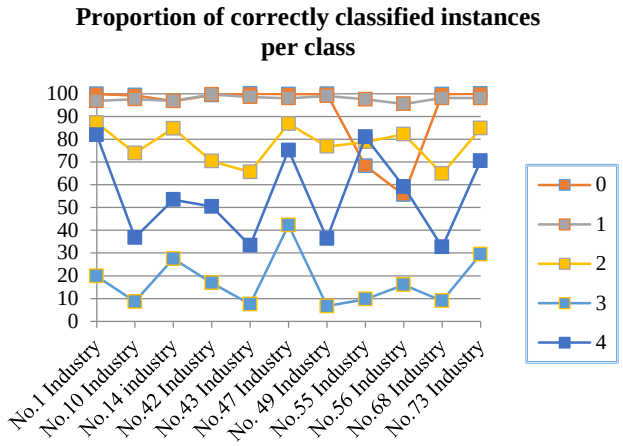

Fig. 5. Proportion of Correctly Classified Instances of the Best Algorithm per Class.

In the case of ROA prediction problem, the random forest algorithm is efficient for classes labeled 0,1,2 with the average proportions of correctly classified instances are 92,59%, 97.78%, and 77.81% relatively. However, it fails for a class labeled 3 with the average correction is only 17.59%, and it reached an average correct proportion at 55.51% for a class labeled 4.

- Result of ROE prediction is presented in Table 10 and Figure 6:

TABLE X. THE PERFORMANCES OF RANDOM FOREST ALGORITHM

Class label	**0**	**1**	**2**	**3**	**4**
No.1 Industry	96.63	87.98	76.44	48.68	87.48
No.10 Industry	94.29	79.07	11.46	45.50	82.40
No.14 industry	92.80	84.99	10.68	45.86	86.99
No.42 Industry	96.19	84.94	10.52	50.68	85.38
No.43 Industry	96.29	82.81	16.07	35.24	86.26
No.47 Industry	96.79	87.84	8.74	56.49	89.37
No. 49 Industry	95.87	85.21	12.89	47.70	82.13
No.55 Industry	95.81	84.42	13.62	28.04	79.41
No.56 Industry	93.01	82.60	13.58	43.52	82.11
No.68 Industry	97.38	79.56	14.34	41.11	86.43
No.73 Industry	95.58	88.33	10.03	57.02	89.14
Average	**96.12**	**85.61**	**18.86**	**49.08**	**86.40**

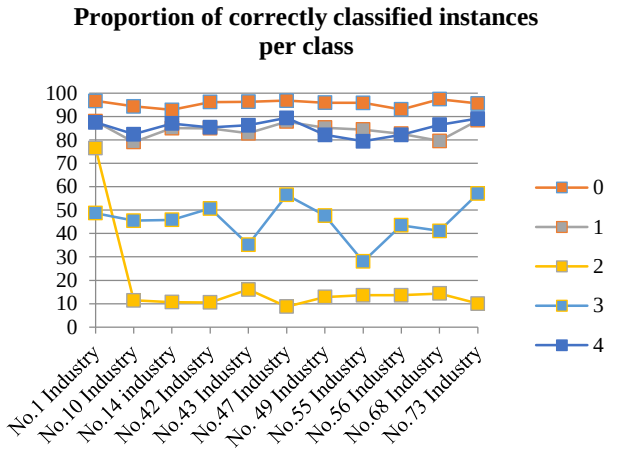

Fig. 6. Proportion of Correctly Classified Instances of the Best Algorithm per Class.

In case of ROE prediction problem, the J48 algorithm is efficient for classes labeled 0,1,4 with the average proportions of correctly classified instances are 96,12%, 85.61%, and 86.40% relatively. However, it fails for class labeled 2 with the average correction is only 18.86%, and it reached the average correct proportion at 49.08% for class labeled 3.

In both cases, the best algorithms are not successful in all classes, however, they succeed in the classification of largest categories including classes labeled 0, 1. Predict these classes are very important to building an early warning system for macroeconomic management as proposing in session 3.

V. CONCLUSION

Forecasting firms' performance for macroeconomic perspective is still an ongoing research area however it is becoming more and more important given the development of digital economy. This study reviewed the disjointed published studies on this research area and consolidated theoretically the application potential of firm performance prediction by ML techniques in macro-economic prediction problem. The study investigated the ability of utilizing micro-level information in macroeconomic monitoring and proposed a framework to process firm-level information to generate on demand information. This study also mentioned the major proven machine learning algorithms for firm's performance prediction used in micro perspectives chronologically, and these algorithms were fundamentals to conduct a research framework that was tested on Vietnam's economic data, keeping in mind that this data contains firm's public information only. This research gave evidence to prove the enormous potential of proposed model for macroeconomic manager.

Particularly, five great machine learning algorithms were studied on data of Vietnamese companies belong to the different industry groups to identify the most suitable model for each industry groups. The applied research methodology had two stages, with the first stage preprocessed and divided the raw data set into sub-datasets belongs to different industry groups and validated these sub-datasets to satisfy the requirements of using machine learning algorithms. The second stage performed main processes including sub-data set partitioning, training, and testing processes for each ML algorithms. The testing result was evaluated by several measurements to ensure comparing comprehensiveness. Besides, this approach opened the ability to improve quality of final models by combination new data dimensional reduction techniques and machine learning algorithms together.

Experiments showed that in both cases, ROA and ROE prediction, regarding the ROC-area the random forest algorithm outperformed all other algorithms and J48 algorithm follows closely. In case of ROA prediction, random forest algorithm was better than J48 according to recall and precision also (table 5). However, J48 was better than random forest according to recall and precision in ROE prediction (table 7). In this study, random forest algorithm was the best for ROA prediction problem and J48 was chosen for ROE prediction problem.

For both ROA and ROE, the best algorithms was not successful in all classes, however, they succeed in classification of the largest categories including classes labeled 0, 1. In fact, predicting these low performance classes is very important to build an early warning system for macroeconomic management.

The proposed approach has high opportunity to use for macroeconomic management because the fast pace of modern economy requires the monitoring decision making in shorter and shorter period time. In fact, the e-government model has been developed and digital economy supplies a large and detail data at firm-level in high frequency and this high-volume data supports to create better machine learning model for firm's performance prediction used in macroeconomic perspective. In expected case, automatic mechanism can be built, and it can generate early warnings for policy makers about economic state.

This study has some limitations on theoretical and experiment sides. The proposed approach utilizes the same selected variables for all machine learning algorithms, and this is not an ideal procedure. Theoretically, each ML algorithm might suit to different variable selection method therefore the ideal procedure should take this fact into account. On the other hand, the testing data set still contains exceptions and bizarre instances. Both limitations are going to be dealt with in the expanded stage of this study in near future.

ACKNOWLEDGMENT

This research is funded by the Ministry of Education and Training and Thuongmai University under grant number B2020-TMA-04.

REFERENCES

[1] J. Gans, S. King, R. Stonecash, and N. G. Mankiw, Principles of economics. Cengage Learning, 2011.

[2] C. N. Wang and V. T. Phan, "Enhancing the accurate of grey prediction for GDP growth rate in Vietnam," Proc. - 2014 Int. Symp. Comput. Consum. Control. IS3C 2014, pp. 1137–1139, 2014, doi: 10.1109/IS3C.2014.295.

[3] J. H. Stock and M. W. Watson, "Forecasting in ation," J. Monet. Econ., vol. 44, no. 2, pp. 293–335, 1999.

[4] A. L. Montgomery, V. Zarnowitz, R. S. Tsay, and G. C. Tiao, "Forecasting the US unemployment rate," J. Am. Stat. Assoc., vol. 93, no. 442, pp. 478–493, 1998.

[5] M. T. Leung, A. Chen, and H. Daouk, "COM&OR_2000_Forecasting Exchange Rates Using General Regression Neural networks.pdf," vol. 27, 2000.

[6] A. Banerjee, M. Marcellinoà, and I. Masten, "Leading Indicators for Euro-area Inflation and GDP Growth*," Oxf. Bull. Econ. Stat., vol. 67, pp. 785–813, 2005.

[7] H. S. Atesoglu, "Growth and fluctuations in the USA: a demand-oriented explanation," Econ. demand led growth, challenging supply-side Vis. long-run, pp. 55–63, 2002.

[8] K. A. Aastveit, K. R. Gerdrup, A. S. Jore, and L. A. Thorsrud, "Nowcasting GDP in real time: A density combination approach," J. Bus. Econ. Stat., vol. 32, no. 1, pp. 48–68, 2014, doi: 10.1080/07350015.2013.844155.

[9] B. Bok, D. Caratelli, D. Giannone, A. M. Sbordone, and A. Tambalotti, "Macroeconomic Nowcasting and Forecasting with Big Data," Annu. Rev. Econom., vol. 10, no. 1, pp. 615–643, 2018, doi: 10.1146/annurev-economics-080217-053214.

[10] J. W. Galbraith and G. Tkacz, "Nowcasting GDP with electronic payments data," 2015.

[11] C. Reinhart, M. Goldstein, and G. Kaminsky, "Assessing financial vulnerability, an early warning system for emerging markets: Introduction," 2000.

[12] A. Ciarlone and G. Trebeschi, "Designing an early warning system for debt crises," Emerg. Mark. Rev., vol. 6, no. 4, pp. 376–395, 2005.

[13] C. Sevim, A. Oztekin, O. Bali, S. Gumus, and E. Guresen, "Developing an early warning system to predict currency crises," Eur. J. Oper. Res., vol. 237, no. 3, pp. 1095–1104, 2014.

[14] J. F. Brito, "Economic Growth as the Result of Firms Aggregate Performance: Evidence from the OECD Countries," Econ. Manag. Res. Proj. An Int. J., vol. 3, no. 1, pp. 24–31, 2013.

[15] K. Martinsen, F. Ravazzolo, and F. Wulfsberg, "Forecasting macroeconomic variables using disaggregate survey data," Int. J. Forecast., vol. 30, no. 1, pp. 65–77, 2014, doi: https://doi.org/10.1016/j.ijforecast.2013.02.003.

[16] P. Fornaro, "Predicting Finnish economic activity using firm-level data," Int. J. Forecast., vol. 32, no. 1, pp. 10–19, 2016, doi: https://doi.org/10.1016/j.ijforecast.2015.04.002.

[17] N. G. . van der Wijst, "The influence of CEO compensation on firm performance and its relation to economic growth," 2018.

[18] E. J. Bartelsman and Z. Wolf, "FORECASTING AGGREGATE PRODUCTIVITY USING INFORMATION FROM FIRM-LEVEL DATA," Rev. Econ. Stat., vol. 96, no. July, pp. 745–755, 2014, doi: 10.1162/REST_a_00395.

[19] K. Dopfer, J. Foster, and J. Potts, "Micro-meso-macro," J. Evol. Econ., vol. 14, no. 3, pp. 263–279, 2004, doi: 10.1007/s00191-004-0193-0.

[20] S. T. Silva, A. A. . Teixeira, and M. R. Silva, "Economics of the firm and economic growth: a hybrid theoretical framework of analysis," J. Organ. Transform. Soc. Chang., vol. 2, no. 3, pp. 255–274, 2005, doi: 10.1386/jots.2.3.255/1.

[21] E. M. Al-Matari, A. K. Al-Swidi, and F. H. B. Fadzil, "The Measurements of Firm Performance's Dimensions," Asian J. Financ. Account., vol. 6, no. 1, p. 24, 2014, doi: 10.5296/ajfa.v6i1.4761.

[22] J. Darroch, "Knowledge management, innovation and firm performance," J. Knowl. Manag., vol. 9, no. 3, pp. 101–115, 2005.

[23] P. Goyal, Z. Rahman, and A. A. Kazmi, "Corporate sustainability performance and firm performance research," Manag. Decis., 2013.

[24] N. Omar, Z. 'Amirah Johari, and M. Smith, "Predicting fraudulent financial reporting using artificial neural network," J. Financ. Crime, vol. 24, no. 2, pp. 362–387, 2017.

[25] C. Demartini, "Performance Management System. A Literature Review," in Performance Management Systems, Springer, 2014, pp. 55–88.

[26] C. Kuzey, A. Uyar, and D. Delen, "The impact of multinationality on firm value: A comparative analysis of machine learning techniques," Decis. Support Syst., vol. 59, no. 1, pp. 127–142, 2014, doi: 10.1016/j.dss.2013.11.001.

[27] X. Jin, J. Wang, T. Chu, and J. Xia, "Knowledge source strategy and enterprise innovation performance: dynamic analysis based on machine learning," Technol. Anal. Strateg. Manag., vol. 30, no. 1, pp. 71–83, 2018, doi: 10.1080/09537325.2017.1286011.

[28] F.-H. Chen and H. Howard, "An alternative model for the analysis of detecting electronic industries earnings management using stepwise regression, random forest, and decision tree," Soft Comput., vol. 20, no. 5, pp. 1945–1960, 2016.

[29] F. Barboza, H. Kimura, and E. Altman, "Machine learning models and bankruptcy prediction," Expert Syst. Appl., vol. 83, pp. 405–417, 2017, doi: 10.1016/j.eswa.2017.04.006.

[30] P. Hajek and R. Henriques, "Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods," Knowledge-Based Syst., vol. 128, pp. 139–152, 2017, doi: 10.1016/j.knosys.2017.05.001.

[31] X. Y. Qiu, P. Srinivasan, and Y. Hu, "Supervised learning models to predict firm performance with annual reports: An empirical study," J. Assoc. Inf. Sci. Technol., vol. 65, no. 2, pp. 400–413, 2014.

[32] Y. Li, W. Jiang, L. Yang, and T. Wu, "On neural networks and learning systems for business computing," Neurocomputing, vol. 275, pp. 1150–1159, 2018, doi: 10.1016/j.neucom.2017.09.054.

[33] Q. Cao, M. A. Thompson, and Y. Yu, "Sentiment analysis in decision sciences research: An illustration to IT governance," Decis. Support Syst., vol. 54, no. 2, pp. 1010–1015, 2013, doi: 10.1016/j.dss.2012.10.026.

[34] P. Hajek, V. Olej, and R. Myskova, "Forecasting corporate financial performance using sentiment in annual reports for stakeholders' decision-making," Technol. Econ. Dev. Econ., vol. 20, no. 4, pp. 721–738, 2014, doi: 10.3846/20294913.2014.979456.

[35] J. Li, H. Bu, and J. Wu, "Sentiment-aware stock market prediction: A deep learning method," 14th Int. Conf. Serv. Syst. Serv. Manag. ICSSSM 2017 - Proc., 2017, doi: 10.1109/ICSSSM.2017.7996306.

[36] D. Delen, C. Kuzey, and A. Uyar, "Measuring firm performance using financial ratios: A decision tree approach," Expert Syst. Appl., vol. 40, no. 10, pp. 3970–3983, 2013, doi: 10.1016/j.eswa.2013.01.012.

[37] E. Zibanezhad, D. Foroghi, and A. Monadjemi, "Applying decision tree to predict bankruptcy," Proc. - 2011 IEEE Int. Conf. Comput. Sci. Autom. Eng. CSAE 2011, vol. 4, pp. 165–169, 2011, doi: 10.1109/CSAE.2011.5952826.

[38] P. du Jardin, "Dynamics of firm financial evolution and bankruptcy prediction," Expert Syst. Appl., vol. 75, pp. 25–43, 2017, doi: 10.1016/j.eswa.2017.01.016.

[39] A. Gepp and K. Kumar, "Predicting Financial Distress: A Comparison of Survival Analysis and Decision Tree Techniques," Procedia Comput. Sci., vol. 54, pp. 396–404, 2015, doi: 10.1016/j.procs.2015.06.046.

[40] E. Basti, C. Kuzey, and D. Delen, "Analyzing initial public offerings' short-term performance using decision trees and SVMs," Decis. Support Syst., vol. 73, pp. 15–27, 2015, doi: 10.1016/j.dss.2015.02.011.

[41] B. Yeo and D. Grant, "Predicting service industry performance using decision tree analysis," Int. J. Inf. Manage., vol. 38, no. 1, pp. 288–300, 2018, doi: 10.1016/j.ijinfomgt.2017.10.002.

[42] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995.

[43] A. Fan and M. Palaniswami, "Selecting bankruptcy predictors using a support vector machine approach," in Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, 2000, vol. 6, pp. 354–359.

[44] K. Kim, "Financial time series forecasting using support vector machines," Neurocomputing, vol. 55, no. 1–2, pp. 307–319, 2003.

[45] A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi, and O. K. Hussain, "Support vector regression with chaos-based firefly algorithm for stock market price forecasting," Appl. Soft Comput., vol. 13, no. 2, pp. 947–958, 2013.

[46] F. E. H. Tay and L. Cao, "Application of support vector machines in financial time series forecasting," Omega, vol. 29, no. 4, pp. 309–317, 2001.

[47] J. H. Min and Y. C. Lee, "Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters," Expert Syst. Appl., vol. 28, no. 4, pp. 603–614, 2005, doi: 10.1016/j.eswa.2004.12.008.

[48] C.-H. Wu, G.-H. Tzeng, Y.-J. Goo, and W.-C. Fang, "A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy," Expert Syst. Appl., vol. 32, no. 2, pp. 397–408, 2007.

[49] F. Lin, C.-C. Yeh, and M.-Y. Lee, "The use of hybrid manifold learning and support vector machines in the prediction of business failure," Knowledge-Based Syst., vol. 24, no. 1, pp. 95–101, 2011.

[50] F. Ecer, "Comparing the bank failure prediction performance of neural networks and support vector machines: The Turkish case," Econ. Res. istraživanja, vol. 26, no. 3, pp. 81–98, 2013.

[51] X. Wu et al., "ERMiner: Sequential rule mining using equivalence classes," Indian J. Sci. Technol., vol. 7, no. 1, pp. 68–76, 2014, doi: 10.1109/ICDM.2004.10117.

[52] M. D. Odom and R. Sharda, "A neural network model for bankruptcy prediction," in 1990 IJCNN International Joint Conference on neural networks, 1990, pp. 163–168.

[53] K. Y. Tam, "Neural network models and the prediction of bank bankruptcy," Omega, vol. 19, no. 5, pp. 429–445, 1991.

[54] P. K. Coats and L. F. Fant, "Recognizing financial distress patterns using a neural network tool," Financ. Manag., pp. 142–155, 1993.

[55] T. B. Bell, "Neural nets or the logit model? A comparison of each model's ability to predict commercial bank failures," Intell. Syst. Accounting, Financ. Manag., vol. 6, no. 3, pp. 249–264, 1997.

[56] D. Vlachos, "Neuro-fuzzy modeling in bankruptcy prediction," Yugosl. J. Oper. Res., vol. 13, no. 2, 2016.

[57] A. Tsakonas, G. Dounias, M. Doumpos, and C. Zopounidis, "Bankruptcy prediction with neural logic networks by means of grammar-guided genetic programming," Expert Syst. Appl., vol. 30, no. 3, pp. 449–461, 2006.

[58] B. J. Zaini, S. M. Shamsuddin, and S. H. Jaaman, "Comparison between rough set theory and logistic regression for classifying firm's performance," J. Qual. Meas. Anal. JQMA, vol. 4, no. 1, pp. 141–153, 2008.

[59] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," arXiv Prepr. arXiv1302.4964, 2013.

[60] J. R. Quinlan, "Simplifying decision trees," Int. J. Man. Mach. Stud., vol. 27, no. 3, pp. 221–234, 1987.

[61] J. H. Stock and M. W. Watson, "Diffusion indexes," NBER Work. Pap., no. w6702, 1998.

[62] J. H. Stock and M. W. Watson, "Macroeconomic forecasting using diffusion indexes," J. Bus. Econ. Stat., vol. 20, no. 2, pp. 147–162, 2002.

[63] E. S. Harris, Tracking the economy with the purchasing managers index. Federal Reserve Bank, 1991.