JOIN (Jurnal Online Informatika)
Volume 5 No. 2 | December 2020: 185-194
DOI: 10.15575/join. join.v5i2.646
p-ISSN: 2528-1682
e-ISSN: 2527-9165

Detection of Fraudulent Financial Statement based on Ratio Analysis in Indonesia Banking using Support Vector Machine

Yuliant Sibaroni[1], Muhammad Novario Ekaputra[2], Sri Suryani Prasetiyowati[3]
[1,2,3] Faculty of Informatics, Telkom University

Article Info

Article history:

Received October 01, 2020
Revised November 12, 2020
Accepted November 27, 2020
Published December 30, 2020

Keywords:

Classification
Feature
Financial Ratio
Fraudulent
Ratio Analysis

ABSTRACT

This study proposes the use of ratio analysis-based features combined with the SVM classifier to identify fraudulent financial statements. The detection method used in this study applies a data mining classification approach. This method is expected to replace the expert in forensic accounting in identifying fraudulent financial statements that are usually done manually. The experimental results show that the proposed classifier model and ratio analysis-based features provide more than 90% accuracy results where the optimal number of features based on ratio analysis is 5 features, namely Capital Adequacy Ratio (CAR), (ANPB) to total earning assets and non-earning assets (ANP), Impairment provision on earning assets (CKPN) to earning assets, Return on Asset (ROA), and Return on Equity (ROE). The contribution of the study is to complement the research of fraudulent financial statements detection, where the classifier method used here is different compare to other research. The selection of banking cases in Indonesia is also unique in this research, which distinguishes it from other research because the financial reporting standards in each country can be different.

Corresponding Author:

Yuliant Sibaroni,
Faculty of Informatics,
Telkom University,
Indonesia
Email: yuliant@telkomuniversity.ac.id

1. INTRODUCTION

The accounting process of a company will produce a product called a financial statement. This statement contains all information regarding the financial position, performance, and changes in the company's financial position in a given period will be presented, and the benefits are intended for a large number of users in making economic decisions and others [1]. A financial statement must be prepared properly based on Financial Accounting Standards (SAK) set by the Indonesian Accounting Association (IAI). The information presented must be factual, objective and not mislead the users, but in some cases, financial statements present incorrect information so it can mislead users in making decisions. This is because there is an opportunity for management to falsify data, using various accounting tactics in managing financial statements so that the statements presented look better before they are announced to public users[2].

The fraudulent of financial statements detection is not an easy thing for ordinary people to do. A special analysis is needed from an expert in forensic accounting to identify data trends, patterns, and unreasonable anomalies from financial statements [3]. A detection system for fraudulent of financial statements is needed because currently many institutions have maintained most of the business transaction data in electronic format with hundreds to billions of transactions while the number of experts in the field of forensic accounting is very small. With the number of these transactions, relying on the ability to investigate manually (without using machines) to uncover suspicious transactions will produce an expensive cost, with respect to time, money, and losses due to human error. In thousands or millions of data, fraudulent activities might remain hidden or not be revealed unless reliable tools are applied to raise the problem towards a solution [4].

In a financial statement, many components are interrelated to one another and lead to certain aspects. In banking financial statements, the relationship is described in interrelated ratios. These interrelated ratios show the relationship between capital formation and financial development. The estuary of this relationship can be expressed by the Capital Adequacy Ratio (CAR). In general, the Capital Adequacy Ratio shows the extent to which banks are exposed to risks that are partially financed by public funds. The risks involved include

credit, statements, securities, invoices. Financial ratio analysis techniques when using small amounts of data are still possible and easy to do. Practitioners usually use the technique based on financial ratio analysis to identify fraudulent financial statements. However, long and comprehensive data series analysis for the entire banking industry can hardly be done manually and separately. Software applications are needed that have the ability to detect fraud in the financial statements through ratio analysis. It is conceivable to do machine learning because a certain pattern will be obtained, which can be used as a system to detect fraudulent financial statements.

Several groups of researchers have conducted research on the fraudulent detection of financial statements. The first approach uses manual analysis based on certain perspectives such as the Fraud Triangle[2], [5], and Diamond Fraud [6]–[8]. The fraud triangle has three elements, namely pressure, opportunity, and rationalization, while the fraud diamond adds one more element, namely, capability. So far, both approaches have yielded good results to identify fraudulent of financial statement detection, but this manual identification process becomes less effective when used to handle large data sizes. The second approach of fraudulent detection of financial statements has been carried out by several researchers using a data mining approach. The data mining-based approach has advantages over the previous approach because it can process large data, and the model obtained can be updated automatically. The data mining method can be implemented by using ratio-based factors to build the model, while the target class is based on the status of the financial statements, namely fake or not. Data mining tasks that are generally used are classification methods, although some are using clustering methods in identifying fraudulent financial statements[9]. Classification methods is a method to identify the category of a new observation based on a labeled training set, while a clustering method is method to identify groups of similar objects based on unlabelled data. Some classification methods that have been used by researchers to detect fraudulent of financial statements are *Logistic Regression* [10], *Naïve Bayes*[11], *Decision Trees*, *Neural Networks* and *Bayesian Belief Networks*[12].

The fraudulent detection of financial statement detection using Logistic Regression conducted by [6] produces a model that has an accuracy rate of about 10% higher using the composition of the optimal feature compared to the same model without feature selection. The fraudulent of financial statements detection is also conducted by [11] using the Naïve Bayes method where the results of the tests in the research on 172 financial statements in China gave quite satisfying results. More complete fraudulent of financial statements detection using data mining classifications is carried out by [12]. In this research, [12] implements and compares 3 classifiers in data mining namely Decision Trees ID3, Neural Networks and Bayesian Belief Networks [8] to detect raudulent of financial statements. The experimental results obtained indicate that Bayesian Belief Networks the method provides the best performance than the Decision Trees ID3 and Neural Networks.

To complement the research on fraudulent detection of financial statements, the Support Vector Machine (SVM) method and Capital Adequacy Ratio (CAR) are used in this research. The use of SVM in this research is based on the fact that this classifier provides good performance in several studies on text classification, and the results are also superior to other classifiers [13]–[15]. The use of SVM method in the identification of fraudulent financial statements is expected to provide better results compared to the methods that have been tried before. The use of the Capital Adequacy Ratio (CAR) in this research is carried out as follows. Each component of the Financial Performance Ratio in a quarterly financial statement will go through the calculation process of increasing the ratio to see changes in the value of these components. Then the results will be used as input features in the SVM classification process. These features are then selected by determining group of features that are most influential and produce the highest level of accuracy. The performance obtained of SVM and CAR methods in this study is the main contribution of this research. Another contribution is the combination of the most optimal ratio-based factors obtained in detecting fraudulent of financial statements.

The rest of the paper is structured as follows. Section 2 describes researches in fraudulent detection of financial statement related to this research. Several classifier methods and the research achievements obtained are discussed in this study. Section 3 explains about evaluating the results of research that has been done. Finally, Section 4 presents conclusions and suggestions for future research.

2. METHOD

The analysis system of classification of financial statements has been developed by researchers. It is just that there has not been much research done for the analysis of banking financial statements. In this study, the plot used also refers to previous studies with additions or modifications, due to adjusting to the existing situations and conditions. Figure 1 explains the system diagram used in this study.

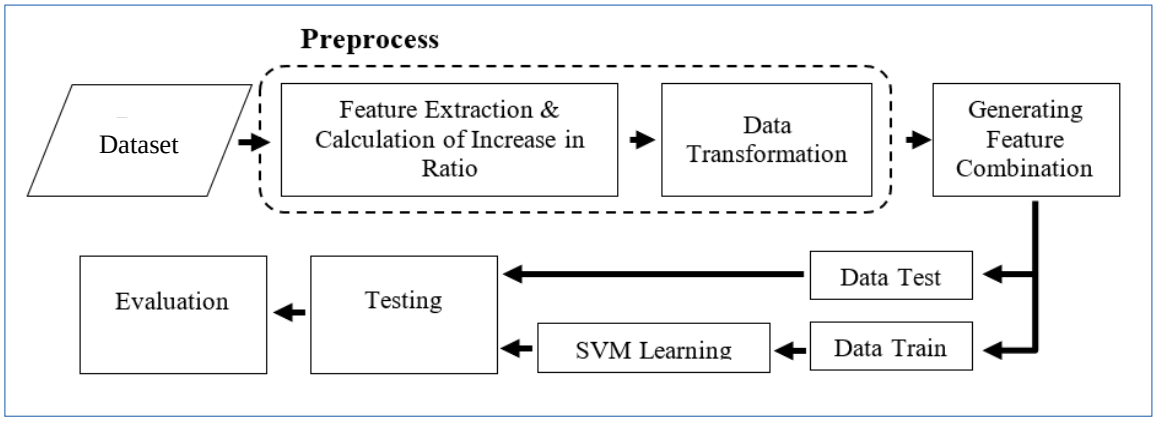

Figure 1. Identification System of Fraudulent Financial Statement

2.1 Dataset

To detect falsification of financial statements, this study uses a dataset of quarterly financial statements from 2015 to 2018 obtained from the Financial Services Authority (OJK). In this case, the author can only use the data provided by the OJK. However, the data used is quite adequate compared to similar studies [12], [16]. This data contains 322 banking financial statements and has two classes, namely problematic and non-problematic, represented by 0 and 1. The problematic and non-problematic classes are 294 and 38 consecutively. Class labeling is done by the Head of Bank Supervision Financial Services Authority (OJK).

2.2 Feature Extraction and Calculation of Increase in Ratio

Feature extraction is the process of extracting important components from financial statements. The final output of this process is a feature matrix. In the data mining approach, datasets in the form of financial reports cannot be processed just like that but must be represented in the form of a feature matrix. The process of extracting important components from large data sets like this makes the feature extraction process known as the process of dimensionality reduction.

In the case of financial statements, reducing the dimensionality of the data cannot be done arbitrarily because in a financial statement, each component is interconnected with one another. Bearing this in mind, this study uses a component of financial performance ratios because these components contain the results of calculating ratios from almost entire components in the financial statements. The financial performance ratio used has ten types of variables, which later will be used as a feature to input the SVM learning process after going through the ratio increase calculation process. Table 1 shows a list of 10 types of financial ratios used in this study, following the provisions of the Financial Services Authority (POJK) Regulation on Commercial Bank Health Rate Assessment, that each bank is required to publish ten types of financial performance ratios.

Table 1. Ten Ratios of Financial Performance Ratio

Symbol	Name of Ratio	Formula
R1	Capital Adequacy Ratio (CAR)	$\frac{Capital}{Risk\ Weighted\ Assets} \times 100\%$
R2	Non-performing earning assets (APB) and non-earning assets (ANPB) to total earning assets and non-earning assets (ANP) (Asset Ratio 1)	$\frac{APB + ANPB}{Total\ Earning\ Aset + ANP} \times 100\%$
R3	Non-performing earning assets (APB) to total earning assets (Asset Ratio 2)	$\frac{APB}{Total\ Earning\ Aset} \times 100\%$
R4	Impairment provision on earning assets (CKPN) to earning assets	$\frac{CKPN}{Earning\ Assets} \times 100\%$
R5	NPL (Non-performing Loan) Gross	$\frac{NPL\ Gross}{Total\ Credit} \times 100\%$
R6	NPL (Non-performing Loan) Net	$\frac{NPL\ Gross - CKPN - Collateral}{Total\ Credit} \times 100\%$
R7	Return on Asset (ROA)	$\frac{EBITDA}{Total\ Asset} \times 100\%$
R8	Return on Equity (ROE)	$\frac{EAT}{Total\ Capital} \times 100\%$
R9	Net Interest Margin (NIM)	$\frac{Net\ Interest\ Income}{Total\ Third\ Party\ Funds} \times 100\%$
R10	Operating Expenses to Operating Revenues (BOPO)	$\frac{Operating\ Expenses}{Operating\ Revenues} \times 100\%$
	Note : Earnings before interest, taxes, depreciation, and amortization (EBITDA) Earnings after Taxes (EAT)	

Anomalies from financial ratio data can be detected through changes in financial ratios, including an increase or decrease in the ratio. For example, regarding Financial Services Authority Regulations (POJK) on Commercial Bank Health Assessment, ROE and ROA ratios are equivalent ratios, meaning that if ROE experiences an increase, ROA will also increase. Suppose it is identified in the data that the ROA ratio has increased while ROE has decreased. In that case, it can be concluded that there are unusual anomalies in the financial ratio. It did not eliminate the possibility of data manipulation has been done that resulted in the ratio of ROA, and ROE is not in reasonable condition. This case can be categorized as a potential forgery. In this study, it is proposed to calculate the annual increase ratio using quarterly financial ratio data. Calculation of an increase in ratio serves to see changes in a financial ratio in a period of one year. The calculation is carried out using the formula (1) as follows

$$Q_1 = \frac{ratio\ (A)\ month\ 6 - ratio\ (A)\ month\ 3}{ratio\ (A)\ month\ 3}$$

$$Q_2 = \frac{ratio\ (A)\ month\ 9 - ratio\ (A)\ month\ 6}{ratio\ A\ month\ 6} \quad (1)$$

$$Q_3 = \frac{ratio\ (A)\ month\ 12 - ratio\ (A)\ month\ 9}{ratio\ (A)\ month\ 9}$$

$$Increase\ Ratio\ (A) = \frac{Q_1 + Q_2 + Q_3}{3}$$

Based on the formula above, A is a variable for the name of the ratio component to be calculated. Because the dataset used is quarterly, the calculation will be done in three quartiles, namely Q1, Q2, Q3. After getting a number from each quartile, the average of each quartile will be calculated, and the results will be the output of this process. This calculation applies to all components of the financial ratio. For example, suppose we have data, as shown in Table 2. The example of calculating the increase in the ratio is as follows:

Table 2. Example of the 2015 NPL ratio dataset

Bank ID	Year	NPL Ratio			
		Month 3	Month 6	Month 9	Month 10
002	2015	20.08234	20.41287	20.59000	20.59000

$$Q_1 = \frac{ratio\ (NPL)\ month\ 6 - ratio\ (NPL)\ month\ 3}{ratio\ (NPL)\ month\ 3} = \frac{20.41287 - 20.08234}{20.08234} = 0.01645$$

$$Q_2 = \frac{ratio\ (NPL)\ month\ 9 - ratio\ (NPL)\ month\ 6}{ratio\ A\ month\ 6} = \frac{20.59000 - 20.41287}{20.41287} = 0.00867$$

$$Q_3 = \frac{ratio\ (NPL)\ month\ 12 - ratio\ (NPL)\ month\ 9}{ratio\ (NPL)\ month\ 9} = \frac{20.59000 - 20.59000}{20.59000} = 0$$

$$Increase\ Ratio\ (NPL) = \frac{Q_1 + Q_2 + Q_3}{3} = \frac{0.01645 + 0.00867 + 0}{3} = 0.00837$$

2.3 Data Transformation

In this stage, the data transformation process will be carried out. The data used in this process is the ratio of financial performance that has gone through the extraction process and calculating the increase in the ratio. The financial ratio data used originally in vertical form because it is the original form of financial statement data. Table 3 illustrates the shape of the data before the transformation process

Table 3. Raw Financial Statements Data

Bank_ID	Year	Statement_ID	Component_Name	Code	Nominal
002	2015	27	Increase Ratio (CAR)	R1	0.0083
002	2015	27	Increase Ratio (Asset Ratio 1)	R2	0.0115
002	2015	27	Increase Ratio (Asset Ratio 2)	R3	0.0067
002	2015	27	Increase Ratio (CKPN)	R4	-0.008
002	2015	27	Increase Ratio (NPL Gross)	R5	-0.306
002	2015	27	Increase Ratio (NPL Net)	R6	-0.041
002	2015	27	Increase Ratio (ROA)	R7	0.0210
002	2015	27	Increase Ratio (ROE)	R8	0.0006
002	2015	27	Increase Ratio (NIM)	R9	0.0242
002	2015	27	Increase Ratio (BOPO)	R10	-0.00025

After going through the process of data transformation which was originally in the form of a vertical, it will change to a horizontal shape as shown in Table 4.

Table 4. Financial Statement Data After the Transformation Process

Bank_ID	Year	R1	R2	R3	R4	…	R9	R10
002	2015	0.0083	0.0115	0.0067	-0.008	…	0.0242	-0.0002
002	2016	0.0137	0.0842	0.0936	0.0576	…	0.0294	0.0235
002	2017	0.0322	0.0865	0.0916	0.1229	…	-0.005	0.0328
002	2018	0.0176	0.1012	0.0932	0.0752	…	-0.009	0.0010
009	2015	0.0257	0.2058	0.2086	0.2984	…	0.0345	0.0514
009	2016	-0.012	0.0654	0.0424	0.0754	…	0.0010	-0.021
009	2017	-0.019	0.1158	0.0858	0.0472	…	0.0002	-0.006
009	2018	0.065	0.0943	0.1248	0.1289	…	0.0313	0.0156

2.4 Determination of Feature Combinations

The determination of feature combinations is the process of determining groups of features that are most influential in terms of label data. Based on the features in Table 4 above, the most influential features will be determined by making a combination of these features. Using the formula (2), a combination of features in sample *r* will be generated.

$$C(n,r) = \frac{n!}{(r!\,(n-r)!)} \tag{2}$$

Considering the relatively few feature ratios, the combination of all factors is carried out in the experiment to obtain the best feature group and obtain the most powerful features. The results of the number of feature combinations are 1023 combinations generated. The combination will be used as a test scenario to determine which features are most relevant marked with the highest accuracy results. Table 5 shows the results of the combination calculations obtained.

Table 5. Calculation Result for Number of Feature Combinations Using Combination Formulas

No.	*n*	*r*	Scenario	Combination	Combination Example
1	10	1	1	10	R1
2	10	2	20	45	R2, R3
3	10	3	58	120	R1, R2, R5
4	10	4	351	210	R4, R5, R6, R7
5	10	5	632	252	R5, R6, R8, R9, R10
6	10	6	783	210	R2, R3, R4, R7, R9, R10
7	10	7	893	120	R1, R2, R4, R6, R7, R8, R9
8	10	8	993	45	R1, R2, R4, R5, R7, R8, R9, R10
9	10	9	1022	10	R2, R3, R4, R5, R6, R7, R8, R9, R10
10	10	10	1023	1	R1, R2, R3, R4, R5, R6, R7, R8, R9, R10
	Total			1023	

The results of making the combination of features above which will later be used as a scenario for testing. The test will also be followed by cross-validation. The combination that obtains the highest accuracy will be concluded that the features in the combination are the most influential or relevant to the label. When all scenarios have gone through a process of learning and cross-validation, the accuration level for each scenario will be obtained.

The outputs from the feature combination testing process will be evaluated. The accuracy obtained from the feature combination testing process will determine whether a feature combination is good or bad. If the accuracy obtained is relatively low, it can be said that a certain combination of features used is not appropriate, in the sense that the combination has features that have low relevance to the label data. From the results of this accuracy, a combination of features with the highest accuracy will be taken for analysis. After finding the scenario with the highest accuracy results, it can be concluded that the combination consists of features that have the highest relevance to the label data.

2.5 SVM Learning

To detect a potential forgery in financial statements, this study uses the SVM classification model. In this research, we use the linear Kernel function as a hyperplane. This is one of the SVM kernels, which has

good performance, although it is the simplest one. Research with the use of a more varied type of Kernel can be done as future research. This section will explain the illustration of the SVM classification model for the two-class classification problem. In the case of two class classifications, assume a given set of samples is a series of input vectors : $X_i \in R^d$ $(i = 1, 2, \dots, n)$ where n is the amount of data, and the label class is stated as $y_i \in \{-1, +1\}$. Suppose that both classes -1 and +1 are assumed to be completely separated by a hyperplane of dimension d, which is defined by equation (3),

$$w \cdot x + b = 0 \quad (3)$$

Data : x_i is classified as class -1 if it satisfies the equation

$$w \cdot x_i + b \leq -1$$

And also classified as +1 class if

$$w \cdot x_i + b > 1$$

The process of estimating the values of w and b is illustrated by the following example. In the example case of the two-class classification problem, suppose given data that has two features as presented in Table 6.

Table 6. Table for the classification data example

Data	X_1	X_2	class
1	1	-1	+1
2	2	-1	+1
3	2	1	+1
4	1	-1	+1
5	5	-1	-1
6	5	1	-1
7	6	0	-1
8	4	0	-1

After plotting data, the next step is to find the support-vector. It can be seen that from the results of the mapping data, three support vectors are chosen and will be represented as variables S_1, S_2, S_3. Three support vectors are selected, i.e. $S_1 = \begin{pmatrix}2\\1\end{pmatrix}$, $S_2 = \begin{pmatrix}2\\-1\end{pmatrix}$, $S_3 = \begin{pmatrix}4\\0\end{pmatrix}$ and will be used to find class boundaries. Then each support vector will be added with 1 as an input bias. Then the support vector value will change to $\tilde{S}_1 = \begin{pmatrix}2\\1\\1\end{pmatrix}$, $\tilde{S}_2 = \begin{pmatrix}2\\-1\\1\end{pmatrix}$, $\tilde{S}_3 = \begin{pmatrix}4\\0\\1\end{pmatrix}$. When each support vector has been added with biased input, the next step is finding the value of three parameters $\alpha_1, \alpha_2, \alpha_3$ based on the following three linear equations :

$$\begin{aligned}
\alpha_1 \tilde{S}_1 \cdot \tilde{S}_1 + \alpha_2 \tilde{S}_2 \cdot \tilde{S}_1 + \alpha_3 \tilde{S}_3 \cdot \tilde{S}_1 &= +1 \\
\alpha_1 \tilde{S}_1 \cdot \tilde{S}_2 + \alpha_2 \tilde{S}_2 \cdot \tilde{S}_2 + \alpha_3 \tilde{S}_3 \cdot \tilde{S}_2 &= +1 \\
\alpha_1 \tilde{S}_1 \cdot \tilde{S}_3 + \alpha_2 \tilde{S}_2 \cdot \tilde{S}_3 + \alpha_3 \tilde{S}_3 \cdot \tilde{S}_3 &= -1
\end{aligned} \quad (4)$$

By substituted each data x_i into equation (4) , it will produce a linear equation system :

$$\alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix} \cdot \begin{pmatrix}2\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix} \cdot \begin{pmatrix}2\\1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix} \cdot \begin{pmatrix}2\\1\\1\end{pmatrix} = 6\alpha_1 + 4\alpha_2 + 9\alpha_3 = +1$$

$$\alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix} \cdot \begin{pmatrix}2\\1\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix} \cdot \begin{pmatrix}2\\-1\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix} \cdot \begin{pmatrix}2\\1\\1\end{pmatrix} = 4\alpha_1 + 6\alpha_2 + 9\alpha_3 = +1$$

$$\alpha_1 \begin{pmatrix}2\\1\\1\end{pmatrix} \cdot \begin{pmatrix}4\\0\\1\end{pmatrix} + \alpha_2 \begin{pmatrix}2\\-1\\1\end{pmatrix} \cdot \begin{pmatrix}4\\0\\1\end{pmatrix} + \alpha_3 \begin{pmatrix}4\\0\\1\end{pmatrix} \cdot \begin{pmatrix}4\\0\\1\end{pmatrix} = 9\alpha_1 + 9\alpha_2 + 17\alpha_3 = -1$$

Based on the linear equation above, the value for the parameter $\alpha_1, \alpha_2, \alpha_3$ will be obtained through a process of elimination and substitution. The process will produce parameter values α_1 = -3.25, α_2 = -3.25, $\alpha_{3=}$ 3.5. After the three-parameter values are obtained, it will be the input into the hyperplane equation formula (5) which will separate the positive class from the negative class.

$$\widetilde{w} = \sum_i \alpha_i \tilde{S}_i \tag{5}$$

By substituting the value obtained previously, the hyperplane equation calculation is as follows :

$$\widetilde{w} = \alpha_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + \alpha_2 \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + \alpha_3 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = (-3.25) \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + (-3.25) \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix} + (3.5) \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

The vector used in this calculation has been added to the bias value of 1. Then we can calculate the entry in $\widetilde{w}$ as a hyperplane with offset b. Therefore, the separating hyperplane equation between positive and negative classes is $y = wx + b$ with $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and offset b = -3.

2.6 Evaluation

Determining the SVM model with the best combination of factors using accuracy measures. The results of the test will be ranked based on the level of accuracy. After finding the scenario group that has the highest accuracy, the scenario group will be used to predict new data. The accuracy of each prediction in a feature combination will be calculated by summing the correct prediction divided by the sum of all data. The formula of accuracy can be seen in equation 6 :

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

Where :

a. True Positive (TP) : Number of fraudulent statements and identified by the system as fraudulent.
b. False Positive (FP) : Number of non-fraudulent statement and identified by the system as fraudulent.
c. False Negative (FN) : Number of fraudulent statement and identified by the system as non-fraudulent
d. True Negative (TN) : Number of non-fraudulent statement and identified by the system as non-fraudulent

Simpler illustrations for TP, TN, FP, and FN can be seen based on the confusion in Table 7. A confusion matrix will then be produced, which is useful for analyzing the quality of the classification model in recognizing tuples from existing classes.

Table 7. The Confusion matrix

Prediction Value	True Value	
	True	False
True	TP	FP
False	FN	TN

3. RESULTS AND DISCUSSION

After going through the testing process, an analysis of the testing results will be carried out. In this section, we will explain the results of the test in the form of measurement metrics, namely accuracy. In addition to measurement metrics, we will also explain which feature groups have the highest accuracy level.

3.1. Result

In this study, testing has been conducted which aims to determine what features are most influential in detecting potential fraud in the financial statements. This testing phase uses a combination of features and cross-validation. In the two-class classification problem, there is a condition where the number of labels of each class is not balanced (Imbalance Class). If you use the train_test_split function, which is a function of dividing test data and training data based on certain ratios, the data will be divided into two parts. In the case of the imbalance class, it is possible that the contents of the split train-data only have the class labeled 0 while the contents of the split test- data have the class labeled both. This condition results in the model being built to

be excessively good in predicting only one class, while the accuracy of the prediction for other classes will be smaller. To overcome this, we need the K-Fold Cross-Validation method. This method will break the data into several parts as many as K, and each iteration will experience a change of test data. Accuracy will also be calculated for each iteration, then the average accuracy of each iteration will be taken.

Cross-validation is an evaluation method commonly used by researchers to measure the performance of classification systems. This method ensures that the level of accuracy obtained in the experiment is the actual performance value because it is measured using the complete data. In addition to using the K-Fold Cross-Validation method, the feature combination selection method can also be used to improve accuracy. This method must be done properly, if not then it will result in a decrease in accuracy, due to the elimination of features that have high relevance to the class label. To overcome this, an analysis must be done to determine what features have high relevance to the class label. By determining the combination of features, it will be known which combination of features produces the highest accuracy.

The scenario will go through a testing process using SVM training and K-Fold Cross Validation, where K=5. The output of the testing process is the accuracy of each scenario. Later, the scenario will be ranked based on its level of accuracy.

Figure 2 shows a graph of the best accuracy obtained for each group of feature combinations. Based on Figure 2, it can be concluded that the optimal number of features is 5 features. The use of the number of features from 1 to 5 features still shows a positive trend. Figure 2 also shows that the use of 6 or more features is no longer effective because no further increase in accuracy can be obtained.

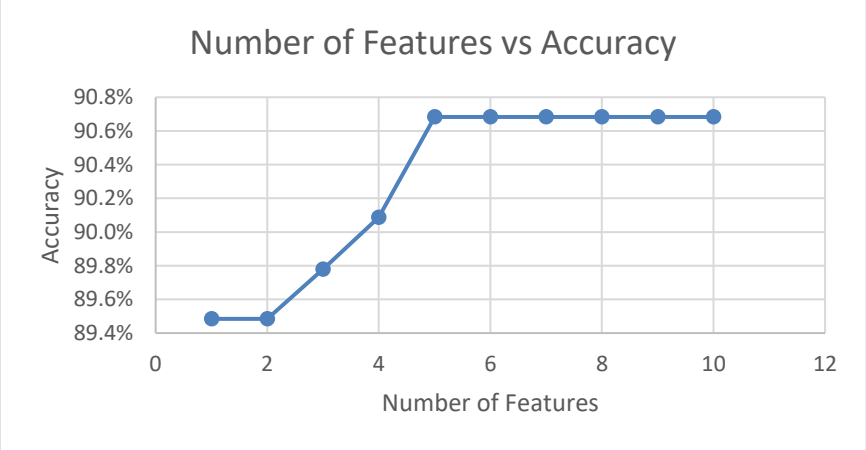

Figure 2. The effect of the number of features on the accuracy obtained

Table 8 shows the features of each group of tested feature combinations from the best accuration of 1 feature to the best accuration of 5 features. Based on table 8, the 5 best features for detecting falsification of financial statements are R1, R2, R4, R7, R8 features namely Capital Adequacy Ratio (CAR), (ANPB) to total earning assets and non-earning assets (ANP), Impairment provision on earning assets (CKPN) to earning assets, Return on Asset (ROA), and Return on Equity (ROE).

Table 8. The scenario ranking for accuracy is based on testing

Number of Features	Features
1	R2
2	R2, R7
3	R2, R7, R8
4	R1, R2, R7, R8
5	R1, R2, R4, R7, R8

3.2. Discussion

Based on Table 8, it can be concluded that the scenario with an accuracy of 90.7 % gets the highest rank among the entire scenarios. The best accuracy value is obtained by using 5 factors, namely Capital Adequacy Ratio (CAR), (ANPB) to total earning assets and non-earning assets (ANP), Impairment provision on earning assets (CKPN) to earning assets, Because in a financial statement the relationship of each component can be described in the form of interrelated ratios that are interrelated to one another and lead to certain aspects. The estuary of this relationship is the capital adequacy ratio (CAR).

From the results obtained, it can be concluded that the best accuracy obtained is 90.7%. Based on these results, it can be seen that the model still has an error of around 9%. This testing process based on_split function, the split used is 80% for training data and 20% for test data. Based on more in-depth analysis, prediction errors occur for 265th to 331st data.

Table 9. List of wrong predictions in testing

Data	R1	R2	R4	R7	R8	Label	Prediction	Analysis
292	0.014	0.744	0.085	-0.031	-0.012	1	0	False Negative
306	0.022	0.627	0.928	-0.354	-0.371	1	0	False Negative

Based on Table 9 above, it can be seen that the 292nd, and 306th data produced incorrect predictions on the test data. The two predictions are categorized into False Negatives in the sense that the statement should be in the fraudulent class but the prediction results categorize the data is in non-fraudulent class. It can be concluded that it needs further analysis to find out why the model used has errors in predicting the two data.

In deeper financial analysis, the two data are also categorized as problematic. When viewed from the value of the ratio component, ROA and ROE values are negative and CAR is still below 8%. Following the Financial Services Authority Regulation (POJK) concerning the Minimum Capital Requirement for Commercial Banks Article 2, Paragraph 3A, the data is categorized as a bank under intensive supervision. However, the first bank's degree of problems is worse than the second bank because the value of the CKPN ratio is very small. It can be concluded that both of them are no longer able to cover the risk of their problem loans. There may be differences in status between the first and second banks, for example, "Intensive 1" and "Intensive 2". According to the POJK, an increase in status could have happened to "the Bank under special supervision".

4. CONCLUSION

This research is motivated by the importance of identifying potential fraud in a financial statement. This identification process uses a data mining classification approach. The accuracy of the classification process is affected by the use of appropriate features and a good classifier model. This research tries to find the most powerful combination of aspect ratio-based features combined with the SVM classifier model. SVM model selection is based on the consideration that this classifier has very good performance in the classification process for various problem domains.

Taking into account the number of features used and considering the highest possible performance, experiments are conducted by testing the performance of all possible combinations of factors. The experimental results show that the highest classification performance is obtained by using a combination of 5 features, namely Capital Adequacy Ratio (CAR), (ANPB) to total earning assets and non-earning assets (ANP), Impairment provision on earning assets (CKPN) to earning assets. The use of 6 or more factors can no longer improve classification performance. Another result obtained is that the CAR factor is the most important factor in the identification process of these fraudulent financial statements. This result does not mean that the other factors are not important, because the combination using other factors also produces high accuracy. The analysis of all factors shows that there is a relationship between these factors.

The highest level of accuracy obtained in the study was 90,7%. Based on these results, it can be seen that the model used still has an error of around 9%. In deeper financial analysis, misclassification of this data is referred to as categorized as problematic. There is contradictory information in these financial statements. The financial statements that failed to be identified are reports that require special analysis from experts. In general, it can be concluded that the use of the proposed financial ratio factor combined with the SVM classifier is effective in identifying fraudulent financial statements automatically. The lack of this research is that not all feature patterns have been identified correctly. For example, when important features have little or negative value, the system does not predict this financial fraud correctly. To deal with this, knowledge-based features of the expert can be developed as the future work of this research.

5. REFERENCES

[1] A. Lako, *Laporan Keuangan & Konflik Kepentingan*. Yogyakarta: Amara Books, 2007.
[2] A. K. Rowland Bismark Fernando Pasaribu, “Fraud Laporan Keuangan Dalam Perspektif Fraud Triangle,” *J. Ris. Akunt. dan Keuang. Fak. Bisnis*, vol. 14, no. 1, 2018.
[3] K. M. Fanning and K. O. Cogger, "Neural network detection of management fraud using published financial data," *Intellegent Syst. Accounting, Financ. Manag.*, 1998.
[4] M. M. Clayton, J. C. Moorman, J. Wilkinson, M. Shackell, and G. Schaffer, "Data Mining : Computer-Adided Forensic Accounting Investigation Techniques," in *A Guide To Forensic Accounting Investigation*, Second Edi., Hoboken: John Wiley & Sons, Inc., 2005, p. 554.
[5] S. Prasmaulida, "Financial Statement Fraud Detection Using Perspective of Fraud Triangle Adopted

By Sas No. 99," *Asia Pacific Fraud J.*, vol. 1, no. 2, p. 317, 2016, doi: 10.21532/apfj.001.16.01.02.24.
[6] A. Arfiyadi and I. Anisykurlillah, "The Detection of Fraudulent Financial Statement with Fraud Diamond Analysis," *Account. Anal. J.*, vol. 5, no. 3, 2016.
[7] N. K. A. Yulistyawati, I. M. S. Suardikha, and I. P. Sudana, "The analysis of the factor that causes fraudulent financial reporting with fraud diamond," *J. Akunt. Audit. Indones.*, vol. 23, no. 1, pp. 1–10, 2019, doi: 10.20885/jaai.vol23.iss1.art1.
[8] M. Yesiariani and I. Rahayu, "Jurnal Akuntansi & Auditing Indonesia Deteksi financial statement fraud : Pengujian dengan fraud diamond," vol. 21, no. 1, 2017.
[9] Q. Deng and G. Mei, "Combining self-organizing map and K-means clustering for detecting fraudulent financial statements," *IEEE Int. Conf. Granul. Comput.*, 2009.
[10] D. Yue, X. Wu, and N. Shen, "Logistic regression for detecting fraudulent financial statement of listed companies in China," *2009 Int. Conf. Artif. Intell. Comput. Intell.*, 2009.
[11] Q. Deng, "Detection of fraudulent financial statements based on Naïve Bayes classifier," *5th Int. Conf. Comput. Sci. Educ.*, 2010.
[12] E. Kirkos, C. Spathis, and Y. Manolopoulos, "Data Mining techniques for the detection of fraudulent financial statements," *Expert Syst. Appl.*, vol. 32, no. 4, pp. 995–1003, 2007, doi: 10.1016/j.eswa.2006.02.016.
[13] M. L. Khodra, D. H. Widyantoro, E. A. Aziz, and B. R. Trilaksono, "Free Model of Sentence Classifier for Automatic Extraction of Topic Sentences," vol. 5, no. 1, pp. 17–34, 2011.
[14] S. Teufel and A. Athar, "Detection of Implicit Citations for Sentiment Detection," *Proc. ACL-12 Work. Discov. Struct. Sch. Discourse, Jeju Island, South Korea, 2012*, no. July, pp. 18–26, 2012.
[15] Y. Sibaroni, D. H. Widyantoro, and M. L. Khodra, "Extend Relation Identification in Scientific Papers Based On Supervised Machine Learning," *2016 Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS 2016)*, 2016.
[16] S. Chen, "Detection of fraudulent financial statements using the hybrid data mining approach," *Springerplus*, vol. 5, no. 1, pp. 1–16, 2016, doi: 10.1186/s40064-016-1707-6.