Classification of Firm External Audit Using Ensemble Support Vector Machine Method

Dewiani[1], Armin Lawi[2], Muhammad Idris Rifai Sarro[3], Firman Aziz[4]
[1,3]Department of Electrical Engineering, Universitas Hasanuddin, Indonesia, 92119
[2]Department of Computer Science, Universitas Hasanuddin, Indonesia, 92119
[4]Faculty of Mathematics and Natural Sciences, Universitas Pancasakti, Indonesia, 90132

{dewiani@unhas.ac.id[1], armin@unhas.ac.id[2], idris30sarro@gmail.com[3], firmanaziz88@gmail.com[4]}

Abstract. Financial fraud is an important problem because it can detrimental firm in the modern business world. An audit is carried out to prevent and be responsible for detecting fraud. External audit is one of the audit practices conducted outside of the firm internal audit by visiting firms in carrying out the work of financial report audit data. The application of machine learning can be used as a solution in the use of data analysis methods needed to solve these problems. This study proposes a Support Vector Machine (SVM) method by combining the Ensemble Bagging model to improve single classification performance. Data comes from 14 different corporate sectors with 777 records. The results showed that the Ensemble Bagging model could improve the accuracy of classification performance from the Support Vector Machine (SVM) method and achieved the highest accuracy of 89.95%. Based on the results of the accuracy obtained, the Support Vector Machine (SVM) method with the Ensemble Bagging model can be used to detect fraud in the firm.

Keywords: *Classification, External Audit, Fraudulent, Support Vector Machine (SVM), Ensemble Bagging.*

1 Introduction

Financial fraud is a crime committed to gain individual or group profits and is an important problem because it can detrimental the firm in modern business world. Audit is an evaluation of the firm carried out by competent, objective and impartial parties, which are referred to as auditors. The aim is to verify that the subject of the audit is completed or running in accordance with the standards, regulations, and practices that have been approved and accepted. Audits can be carried out both internally and externally, internal audits are carried out by firms auditing from various segments and so that the firm's operating procedures run efficiently [1]. Whereas, external audits are carried out for the process of checking financial records and ensuring that the firm's financial statements are compiled following common accounting rules and working principles [2].

ICOST 2019, May 02-03, Makassar, Indonesia

DOI 10.4108/eai.2-5-2019.2284605

The Fraud Detection System (FDS) should not only detect fraud cases efficiently, but also save costs in the sense that the costs invested in screening transactions should not be higher than the losses due to fraud that occur [3]. In minimizing costs in audit practices to detect financial fraud, it is important to use statistical data analysis methods such as machine learning.

Machine Learning has been applied in many aspects of computational financial analysis [4]. Some financial areas where machine learning has been applied include: prediction of bankruptcy [5][6], credit card fraud [3][7], loan decisions [8], financial fraud [9], etc. However, this research is related to the use of machine learning, which is to detect a firm financial fraud because it is one of the most interesting and important issues for the company to conduct a more in-depth study.

Several applications of machine learning that have been applied in detecting fraud, including research [2] using Bayes Net and J48 because they provide stable results in testing K-Fold validation as the feasibility of classification for conducting risk assessments that are efficient in determining decisions on audit field work because accuracy of up to 93%. Research [10] uses three classification methods: Decision Tree (DT), Logistic Regression (LR), and Artificial Neural Network (ANN) to carry out the fraud detection classification process, although the three methods are good at classification, but ANN shows more performance both of the other two classification methods. Research [11] uses and compares three classification methods: Support Vector Machine (SVM), Decision Trees (DT), and Neural Network (NN) to predict auditor choices, based on evaluation of 10-Fold cross validation, Decision Tree outperforms two other classification methods , achieving an average accuracy rate of 83.73%. In addition, the classification method to provide good performance enhancements can use an ensemble model based on some training to solve the same problem and then the output of a single classification is combined into one classifier [12], [13], [14].

In this reasearch, the classification in detecting firm fraud was proposed to use the Ensemble Bagging model with the Support Vector Machine (SVM) method. By implementing Ensemble Bagging in classification using the Support Vector Machine (SVM) method, it is expected to provide an increase in the level of accuracy in classification.

2 Material and Methods

External audit is a problem in classification to determine decisions regarding financial statements between two classes of claims "legitimate" and "fraud" [15]. Classification results are an important process in assisting auditors in firm management. Therefore, external audits are very important to be implemented for firm to minimize the risk of financial fraud.

a. Proposed Framework

The work flow of the complete audit modeling process will be presented in Fig 1. The results of the proposed framework can help to detect financial fraud in the firm and provide a basic description of decision making for firm risk assessment during audit planning.

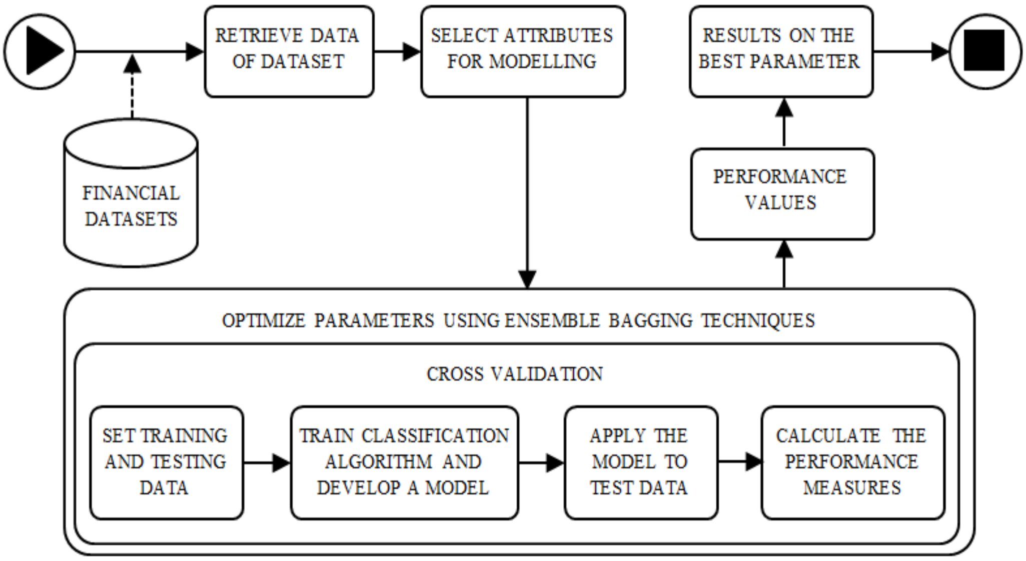
RETRIEVE DATA OF DATASET
SELECT ATTRIBUTES FOR MODELLING
RESULTS ON THE BEST PARAMETER
FINANCIAL DATASETS
PERFORMANCE VALUES
OPTIMIZE PARAMETERS USING ENSEMBLE BAGGING TECHNIQUES
CROSS VALIDATION
SET TRAINING AND TESTING DATA
TRAIN CLASSIFICATION ALGORITHM AND DEVELOP A MODEL
APPLY THE MODEL TO TEST DATA
CALCULATE THE PERFORMANCE MEASURES

Fig.1. The Work Flow of Modeling Process

b. Classification Using Support Vector Machine (SVM)

Support Vector Machine: SVMs are the most widely used technique for classification of different type of type datasets. It searches for data in a space i.e. boundary between two classes that are present at the edge of an area and pass them as support vectors. It is a preferred technique for classification [16]. SVM is fundamentally a classification technique. The system is trained to determine a decision boundary between classes of "legitimate" and "fraudulent" claims. Then each claim is compared with that decision boundary and is placed into either legitimate or fraudulent class [17].

The concept of the SVM diagram is shown in Fig. 2. Processing of SVM is divided into linear divisibility and linear non-divisibility.

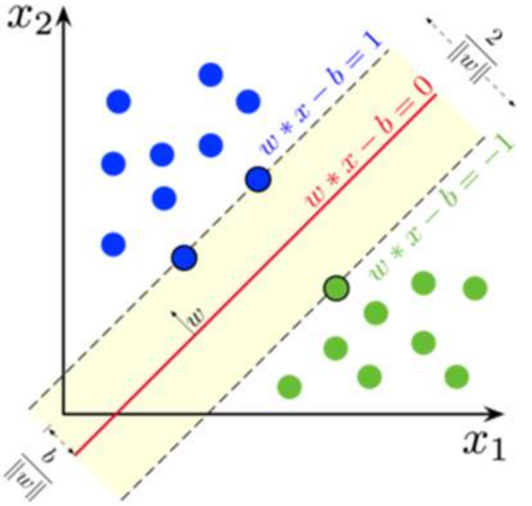

Fig.2. Support Vector Machine (SVM) Concept Diagram

Linear Divisibility

Let the training set be $(x_1, y_1), ..., (x_l, y_l), x \in R^n, y \in \{1, -1\}$, in which x is the input variable, and the data will classify a super-plane into two categories, one is 1, and the other is −1. If such data can be distinguished correctly, and the nearest vector of each classification is at a maximum distance from the super plane, then the super plane is most suitable for divisibility. We express the super plane with the following formula:

$$y_i\ (w,x) + b \geq 1,\ i = 1,...,l \tag{1}$$

The maximum spacing of the two nearest vectors of Eq. (1) is as shown in Eq. (2) :

$$\begin{aligned} p(w,b) &= \min_{x_i y_i = -1} d(w,b:\ x_i) + \min_{x_i y_i = 1} d(w,b:\ x_i) \\ &= \min_{x_i y_i = -1} \frac{|(w,x_i)+b|}{||w||} + \min_{x_i y_i = 1} \frac{|(w,x_i)+b|}{||w||} = \frac{2}{||w||} \end{aligned} \tag{2}$$

The maximum distance (2) can be expressed as minimum $\phi(w) = \frac{1}{2}\|w\|^2$, Lagrange relaxation is used to calculate the super plane most suitable for divisibility, as in Eq. (3) :

$$w^* = \sum_{i=1}^{l} a_i y_i x_i \qquad b^* = -\frac{1}{2}(w^*, x_r + x_s) \tag{3}$$

Where x_r, x_s are any support vectors that each classification can meet $\alpha_s > 0$, $y_r = -1$, $y_s = 1$. We acquire the clear hard classifier, as in Eq. (4) :

$$f(x) = sgn\ ((w^*,\ x) + b) \tag{4}$$

If the condition of incapable or complete classification is not considered, the soft classifier is as shown in Eq. (5) :

$$f(x) = h(w^*, x) + b \quad \text{where} \quad h(z) = \begin{cases} -1: & z < -1 \\ z: & -1 \leq z \leq 1 \\ +1: & z > 1 \end{cases} \tag{5}$$

Linear Non-Divisibility

Imports the concept of the cost function as related to wrong classifications in order to calculate the super plane most suitable for divisibility, which can be expressed as Eq. (6) :

$$\begin{gathered} \min \phi(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ s.t\ \ y_i[(w,x) + b] \geq 1 - \xi_i,\ \ i = 1, \ldots, l \end{gathered} \tag{6}$$

Where ξ_i is the error item of wrong classification, C is a given parameter value, and Lagrange relaxation can be obtained, as in Eq. (7) :

$$\phi(w,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i - \sum_{i=1}^{l}\alpha_i(y_i[w^T x_i + b] - 1 + \xi_i) - \sum_{i=1}^{l}\beta_i\xi_i \quad (7)$$

Where α, β are Lagrange constants.

c. Ensemble Bagging

The ensemble method can reduce classification errors effectively, and is believed to perform well compared to the use of a single classifier. The main idea of the ensemble method is to combine several sets of models that solve a similar problem to obtain a more accurate model [14]. Compared to an individual classifier, they only learn and train a set of data only. But ensemble classifiers learn and train the various data generated from the original dataset and the results will build a set of hypotheses from the data trained and produce better accuracy.

Several Ensemble classification techniques have been developed such as Bagging, Boosting, Random Forest, and Rotation Forest. However, this study only focuses on Ensemble Bagging classification techniques because the algorithm used is the simplest but can overcome instability in complex models with relatively smaller data sets.

Bagging (Bootstrap Aggregating) algorithm creates M bootstrap samples T_1, T_2,. . . , T_M randomly drawn from the original training set T of size n. Each bootstrap sample T_i of size n is then used to train a base classifier C_i. Predictions on new observations are made by taking the majority vote of the ensemble C^* built from C_1, C_2,…, C_M[18]. Works of the ensemble bagging algorithm is given in Fig.3.

```
Algorithm for ensemble Bagging:
Given training set of size $n$ and base classification algorithm $C_t(x)$.
1. Input sequence of training samples $(x_1 : y_1)$, .... $(x_n : y_n)$ with labels $y \in Y = (-1, 1)$.
2. Initialize propability  for each example in learning set $D_1(i) = \frac{1}{n}$ and set $t = 1$.
3. Loop while $t < B = 100$ ensemble members :
   a. Form training set of size $n$ by sampling with replacement from distribution $D_t$.
   b. Get hyphothesis $h_t : X \rightarrow Y$.
   c. Set $t = t + 1$
   End of loop
4. Output the final ensemble hyphothesis
   $C^*(x_i) = h_{final}(x_1) = argmax \sum_{t=1}^{B} I(C_t(x) = y)$
```

Fig.3. The Ensemble Bagging Algorithm

3 Results and Discussion

a. Data Set

In this research, the data used to evaluate the accuracy of the proposed model is data sourced from the General Office Auditor (AGO). Data from 14 different corporate sectors totals 777 records. Information about the firm sectors can be seen in the Table. 1.

Table 1. Sektor Sasaran Perusahaan

Sector ID	Information of Target Sector
1	Agriculture (AG)
2	Animal Husbandry (AH)
3	Buildings and Roads (BR)
4	Communication (CM)
5	Corporate (CO)
6	Electrical (EL)
7	Fisheries (FS)
8	Forest (FR)
9	Industries (ID)
10	Irrigation (IR)
11	Land (LN)
12	Public Health (PH)
13	Science and Technology (ST)
14	Tourism (TO)

b. Performance Evaluation

A number of performance metrics can be used to report classification performance in detecting fraud including Confusion Matrix, Sensitivity, Specificity, False Positive Level, Balanced Classification Level and Matthews Correlation Coefisient. In the matter of detecting fraud, Confusion Matrix is used to calculate misclassified marks and is usually referred to as a contingency table as seen in table 2.

Table 2. The Confusion Matrix

Actual / Prediction	Predicted	
	Positive	Negative
Legitimate	True Positive (TP)	False Positive (FP)
Fraudulent	False Negative (FN)	True Negative (TN)

Based on the table above, it shows that positive values represent transactions that are legitimate and negative values represent fraudulent transactions.

Specificity is defined as the number of fraud case predictions to the total number of fraud cases.

$$\text{Specificity} = TN / (TN + FP) \tag{8}$$

Sensitivity is defined as the number of legitimate predictions compared to the total number of legitimate transactions. In fraud detection, the most important measure is specificity or fraud detection rate, as a higher value of recall means a lowest financial loss to the firm.

$$\text{Sensitivity} = TP / (TP + FN) \tag{9}$$

Accuracy gives the overall efficacy of the proposed system. It is defined as the total number of predictions to the total number of cases. The accuracy level of the model can be given with the following equation (10).

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \tag{10}$$

c. Experimental Result

The fraud detection models were trained and tested using Phyton GNU v.3.4.3. We have used K-fold cross validation (k=10) in the process of training and testing the different models. The average performance results are then recorded. This methodological approach ensures that all data were represented once as a test data and several times as a training data producing accurate results.

Table 3. Classification Results

Classification	Ensemble	Accuracy %	Sensitivity %	Specificity %
Support Vector Machine (SVM)	-	87,26	94,70	74,44
	Bagging	89,95	95,70	80,49

Table 3. show the classification results on the Support Vector Machine (SVM) classification model resulting in an accuracy of 87.26%. Furthermore, the results of the proposed model, Ensemble Support Vector Machine (SVM) using the Bagging method resulted in an accuracy of 89.95%. The application of the classification method using the Ensemble Support Vector Machine (SVM) method achieved an increase with a higher level of accuracy in the data set of 2.69%.

4 Conclusions

In this research, proposed an Ensemble Bagging model with Support Vector Machine (SVM) method in order to see the results of performance in terms of the level of accuracy in detecting fraudulent acts that occur in the company through an external audit. Data is collected from 14 different company sectors with 777 records to be implemented in the proposed method. Based on the results obtained, the accuracy possessed by the support vector machine method can be increased to 89,95% by using Ensemble Bagging, so that the Support Vector Machine (SVM) method with Ensemble Bagging model can be used to detect corporate financial fraud.

For future work, the authors expect to improve classification performance through a machine learning approach using different ensemble model to produce better performing classifications.

Acknowledgements

The authors would like to thank the University of Hasanuddin and our colleagues from Panca Sakti University, for their support for this work and also for anonymous reviewers, for their insightful comments that helped review the content in improving the quality of this research paper.

References

[1] M. Jans, N. Lybaert, and K. Vanhoof, "Internal Fraud Risk Reduction : Results of A Data Mining Case Study," *Int. J. Account. Inf. Syst.*, vol. 11, pp. 17–41, 2010. DOI:10.1016/j.accinf.2009.12.004

[2] N. Hooda, S. Bawa, and P. S. Rana, "Fraudulent Firm Classification: A Case Study of an External Audit," *Appl. Artif. Intell. An Int. J.*, vol. 32, no. 1, pp. 48–64, 2018. DOI: 10.1080/08839514.2018.1451032

[3] S. Venkata Suryanarayana, G. N. Balaji, and G. Venkateswara Rao, “Machine Learning Approaches for Credit Card Fraud Detection,” *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 917–920, 2018. DOI: 10.14419/ijet.v7i2.9356

[4] B. Li, J. Yu, J. Zhang, and B. Ke, “Detecting Accounting Frauds in Publicly Traded U . S . Firms : A Machine Learning Approach,” *JMLR Work. Conf. Proc.*, vol. 45, pp. 173–188, 2015.

[5] K. Nagaraj and A. Sridhar, “A Predictive System for Detection of Bankruptcy Using Machine Learning Techniques,” no. 2, 2015.

[6] F. Mai, S. Tian, C. Lee, and L. Ma, “Deep Learning Models for Bankruptcy Prediction Using Textual Disclosures,” *Eur. J. Oper. Res.*, vol. 274, no. 2, pp. 743–758, 2018.

[7] N. Khare and S. Y. Sait, “Credit Card Fraud Detection Using Machine Learning Models and Collating Machine Learning Models,” *Int. J. Pure Appl. Math.*, vol. 118, no. 20, pp. 825–838, 2018.

[8] A. Goyal and R. Kaur, “Accuracy Prediction for Loan Risk Using Machine Learning Models,” *Int. J. Comput. Sci. Trends Technol.*, vol. 4, no. 1, pp. 52–57, 2016.

[9] K. Tangod and G. Kulkarni, “Detection of Financial Statement Fraud and Feature Selection Using Data Mining Techniques,” *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 4, no. 7, pp. 549–555, 2015.

[10] G. Özdağoğlu, A. Özdağoğlu, Y. Gümüş, and G. K.- Gümüş, “The Application of Data Mining Techniques in Manipulated Financial Statement Classification : The Case of Turkey,” *J. AI Data Min.*, vol. 5, no. 1, pp. 67–77, 2017.

[11] E. Kirkos, C. Spathis, and Y. Manolopoulos, “Support Vector Machines , Decision Trees and Neural Networks for Auditor Selection,” *J. Comput. Methods Sci. Eng.*, vol. 8, pp. 213–224, 2008.

[12] 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

[13] M. Zareapoor and P. Shamsolmoali, “Application of Credit Card Fraud Detection : Based on Bagging Ensemble Classifier,” *Procedia Comput. Sci.*, vol. 48, no. C, pp. 679–685, 2015.

[14] G. Tuysuzoglu, D. Birant, and A. Pala, “Ensemble Methods in Environmental Data Mining,” *Sch. Enviromental Sci.*, pp. 1–16, 2018.

[15] A. Bănărescu, “Detecting and Preventing Fraud with Data Analytics,” *Procedia Econ. Financ.*, vol. 32, no. 15, pp. 1827–1836, 2015.

[16] A. Tiwari and N. Hooda, “Machine Learning Framework for Audit Fraud Data Prediction,” *ResearchGate*, vol. 7, no. 6, pp. 164–167, 2018.

[17] V. Rawte and G. Anuradha, “Fraud Detection in Health Insurance Using Data Mining Techniques,” *Int. Conf. Commun. Inf. Comput. Technol. ICCICT*, vol. 978, no. 1, pp. 723–729, 2015.

[18] F. Aziz and A. Lawi, “On Increasing Accuracy of Ensemble Logistics Regression Classifier by Estimating the Newton Raphson Parameter in Credit Scoring,”

ResearchGate, no. 3, pp. 10–14, 2019. DOI: 10.1109/CAIPT.2017.8320700