Fraud Detection using Machine Learning in e-Commerce

Adi Saputra[1], Suharjito[2]
Computer Science Department, BINUS Graduate Program–Master of Computer Science
Bina Nusantara University Jakarta, Indonesia 11480

***Abstract*—The volume of internet users is increasingly causing transactions on e-commerce to increase as well. We observe the quantity of fraud on online transactions is increasing too. Fraud prevention in e-commerce shall be developed using machine learning, this work to analyze the suitable machine learning algorithm, the algorithm to be used is the Decision Tree, Naïve Bayes, Random Forest, and Neural Network. Data to be used is still unbalance. Synthetic Minority Over-sampling Technique (SMOTE) process is to be used to create balance data. Result of evaluation using confusion matrix achieve the highest accuracy of the neural network by 96 percent, random forest is 95 percent, Naïve Bayes is 95 percent, and Decision tree is 91 percent. Synthetic Minority Over-sampling Technique (SMOTE) is able to increase the average of F1-Score from 67.9 percent to 94.5 percent and the average of G-Mean from 73.5 percent to 84.6 percent.**

***Keywords*—Machine learning; random forest; Naïve Bayes; SMOTE; neural network; e-commerce; confusion matrix; G-Mean; F1-score; transaction; fraud**

I. Introduction

Insight of previous research results on internet users in Indonesia as released on October 2019 edition of Marketeers Magazine [1], according to the research the number of internet users in Indonesia on 2019 alone, had reached 132 million users, an increase from the previous year at 143.2 million users show in Fig. 1.

The increasing number of internet users in Indonesia has triggered market players in Indonesia to try opportunities to develop their business through internet media. One method used is to develop an E-Commerce business [3].

Based on statistical data obtained by Statista.com, it is shown that the number of retail e-Commerce (electronic commerce) sales in Indonesia will grow 133.5% to the US $ 16.5 billion or around IDR 219 trillion in 2022 from the position in 2017. This growth is supported by the rapid advances in technology that provide convenience for consumers to shop.

Huge number of transactions in e-commerce raises the potential for new problems namely fraud in e-commerce transactions shows in Fig. 2. The number of e-commerce-related frauds has also increased every year since 1993. As per a 2013 report, 5.65 cents lost due to a fraud of every $ 100 in e-commerce trading turnover. Fraud has reached more than 70 trillion dollars until 2019 [5]. Fraud detection is one way to reduce the amount of fraud that occurs in e-commerce transactions.

Fraud detection that has developed very rapidly is fraud detection on credit cards ranging from fraud detection using machine learning to fraud detection using deep learning [6] but unfortunately fraud detection for transactions on e-commerce is still small, fraud detection research on e-commerce commerce is still not much so far, fraud detection research on e-commerce is only limited to the determination of features or attributes [7] which will be used to determine the nature of fraud or non-fraud transactions in e-commerce.

The dataset used in this paper has a total of 151,112 records, the dataset classified as fraud is 14,151 records, the ratio of fraud data is 0.093 percent. Datasets that have very small ratios result in an imbalance of data. Imbalance data results in accuracy results that are more inclined to majority data than minority data. The dataset used results more in the classification of the majority of non-fraud than fraud. Accuracy results that are more inclined to majority data make the classification results worse; handling imbalance data using the SMOTE (Synthetic Minority Oversampling Technique).

Recent research about fraud detection in e-commerce transactions still determine feature extraction [8], purpose of this paper is to find the best model to detect fraud in e-commerce transactions.

In this paper research fraud transaction in ecommerce, research use dataset from Kaggle, improve classification machine learning using SMOTE, SMOTE using to handling unbalance data, after using SMOTE, dataset will be training using machine learning. Machine learning is decision tree, Naïve Bayes, random forest, and neural network machine learning to determine accuracy, precision, recall, G-mean, F1-Score.

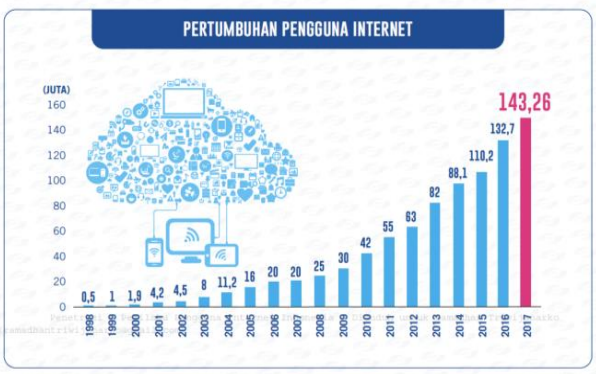

Fig. 1. Growth of Internet users [2].

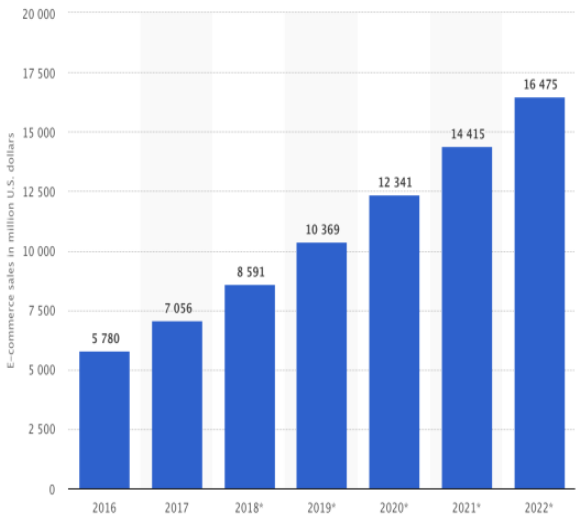

Fig. 2. Sales of e-Commerce, Statista.com [4].

II. Related Works

Fraud detection that has developed very rapidly is fraud on credit cards. Many studies discuss the fraud method. One of the studies carried out using deep learning is auto-encoder and restricted Boltzmann machine [9]. Deep learning is used to build a fraud detection model that runs like a human neural network, where data will be made in several layers that are tiered for the process, starting from the Encoder at layer 1 hinge decoder at layer 4. The researcher compares the deep learning method with other algorithms such as Hidden Markov Model (HMM) [10].

Credit card fraud detection research was also using machine learning [11] machine learning used as a decision tree algorithm, naïve Bayes, neural networks, and random forests.

Decision tree is one algorithm that is widely used in fraud detection because it is easy to use. Decision tree is a prediction model using tree structure or hierarchical structure.

Naïve Bayes is used in fraud detection credit cards because Naïve Bayes is a classification with probability and statistical methods. Naïve Bayes is very fast and quite high inaccuracy in real-world conditions neural network on fraud detection credit cards uses genetic algorithms to determine the number of hidden layer architectures on neural networks [12] with genetic algorithm, the genetic algorithm produces the most optimal number of hidden layers [13]. Fraud detection on credit cards also uses random forest [14]. Random forest uses a combination of each good tree and then combined into one model. Random Forest relies on a random vector value with the same distribution on all trees where each decision tree has a maximum depth [15].

Research on fraud detection in e-commerce is still not much so far. Fraud detection research on e-commerce is only limited to the determination of features or attributes that will be used to determine the nature of the fraud or non-fraud transactions [16]. The study describes the extraction attribute/feature process used to determine behavior in e-commerce transactions. This attribute is used as fraud detection in e-commerce. This attribute determines the transaction conditions.

Another research on fraud detection in e-commerce is a reason transaction based on the attributes or features that exist in e-commerce transactions. The features/attributes used are features of the transaction, namely invalid rating, confirmation interval, average stay time on commodities, a feature of buyer namely real name, positive rating ratio, transaction frequency.

Imbalance of data results in suboptimal classification results. The dataset on the paper has a total number of 151,112 records, the dataset classified as fraud is 14,151 records, and the ratio of fraud data is 0.093 percent. Synthetic Minority Oversampling Technique (SMOTE) is one of the methods used to make data into balance, Synthetic Minority Oversampling Technique (SMOTE) [17] is one of the oversampling methods that work by increasing the number of positive classes through random replication of data, so that the amount of data positive is the same as negative data. The way to use synthetic data is to replicate data in a small class. The SMOTE algorithm works by finding k closest neighbor for a positive class, then constructing duplicate synthetic data as much as the desired percentage between randomly and positively chosen k classes.

Recent paper about fraud detection only limited to the determination of features or attributes. Improvement fraud detection in e-commerce is used machine learning. Machine learning used is the Decision Tree, Naïve Bayes, Random Forest, and Neural Network.

III. Research Methodology

This paper aims to classify e-commerce transactions that include fraud and non-fraud using machine learning, namely Decision Tree, Naïve Bayes, Random Forest, and Neural Network. The research process is carried out as shown Fig. 3.

The classification process begins with the feature selection process in the dataset. After the feature is determined, what is done is preprocessing data using PCA, the process is carried out by transformation, normalization, and scaling of features so that the features obtained can be used for classification after the classification process is done by the SMOTE (Synthetic Minority Oversampling Technique) process. SMOTE is useful for making imbalance data into balance. The SMOTE (Synthetic Minority Oversampling Technique) process is useful for dealing with data imbalance problems in fraud cases, because fraud cases are usually below 1 percent, so as to reduce the majority class in the dataset. The majority class can make the classification more directed to the majority class so that the predictions of the classification are not as expected; the results of the SMOTE dataset transaction fraud process will be balanced [18]

Machine learning used in the classification process is decision tree, random forest, artificial neural network, and naïve Bayes. This machine-learning algorithm will be compared to find the best accuracy results from the transaction dataset in e-commerce.

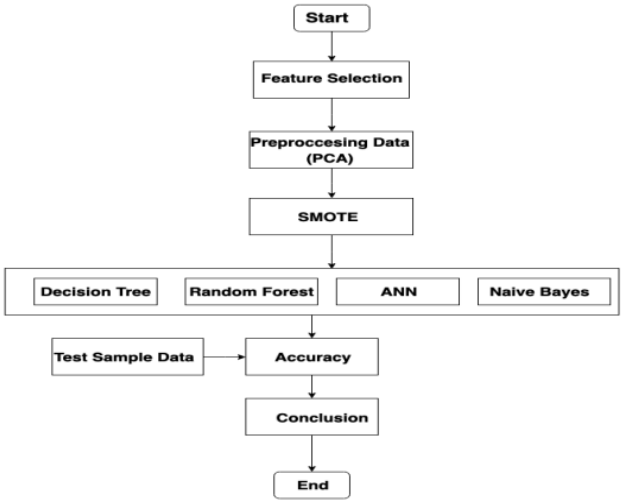

Fig. 3. Research Steps.

A. Preprocessing Data

Preprocessing is used to extract, transform, normalize and scaling new features that will be used in the machine learning algorithm process to be used. Preprocessing is used to convert raw data into quality data. In this study preprocessing uses PCA (Principle Component Analysis) with the features [19] of extraction, transformation, normalization and scaling.

PCA is a linear transformation commonly used in data compression and is a technique commonly used to extract features from data at a high-dimensional scale. PCA can reduce complex data to smaller dimensions to display unknown parts and simplify the structure of data. PCA calculations involve calculations of covariance matrices to minimize reduction and maximize variance.

B. Decision Tree

Decision trees are useful for exploring fraud data, finding hidden relationships between a number of potential input variables and a target variable. Decision tree [20] combines fraud data exploration and modeling, so it is very good as a first step in the modeling process even when used as the final model of several other techniques [21].

Decision tree is a type of supervised learning algorithm; a decision tree is good for classification algorithm. Decision tree divides the dataset into several branching segments based on decision rules, this decision rule is determined by identifying a relationship between input and output attributes.

- Root Node: This represents the entire population or sample, and this is further divided into two or more.
- Splitting: This is the process of dividing a node into two or more sub-nodes.
- Decision Node: When a sub-node is divided into several sub nodes.
- Leaf / Terminal Node: Unspecified nodes are called Leaf or Terminal nodes.
- Pruning: When a sub-node is removed from a decision.
- Branch / Sub-Tree: Subdivisions of all trees are called branches or sub-trees.
- Parent and Child Node: A node, which is divided into sub-nodes [22].

The fraud detection architecture using a decision tree consists of the root node, internal node and leaf node of the decision tree architecture as shown Fig. 4.

C. Naïve Bayes

Naïve Bayes predicts future opportunities based on past experience [23], it uses the calculation formula as below.

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)} \quad (1)$$

Where:

B: Data with unknown classes

A: The data hypothesis is a specific class

P(A|B): Hypothesis probability based on conditions (posterior probability)

P (A): Hypothesis probability (prior probability)

P(B|A): Probability-based on conditions on the hypothesis

P (B): Probability A

By using the formula above can be obtained opportunities from fraud transactions and non-fraud transactions

D. Random Forest

Random forest (RF) is an algorithm used in the classification of large amounts of data. Random Forest (RF) is a development of the Classification and Regression Tree (CART) method by applying the bootstrap aggregating (bagging) method and random feature selection Architecture Random forest as shown in Fig. 5.

Random forest is a combination of each good transaction fraud tree which is then combined into one model. Random Forest relies on a random vector value with the same distribution on all trees, each decision tree in e-commerce fraud detection which has a maximum depth. The class produced from the classification process is chosen from the most classes produced by the decision tree.

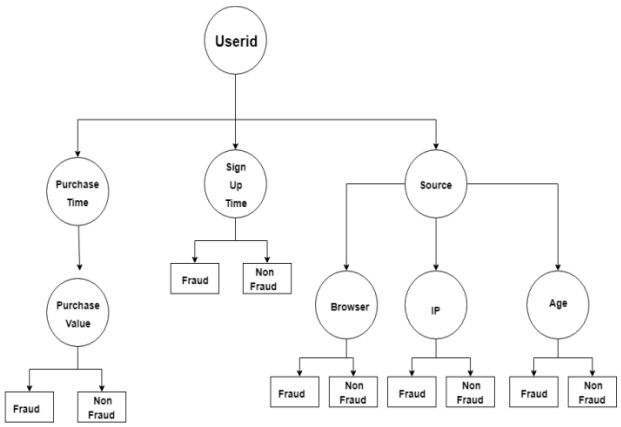

Fig. 4. Architecture of Decision Trees.

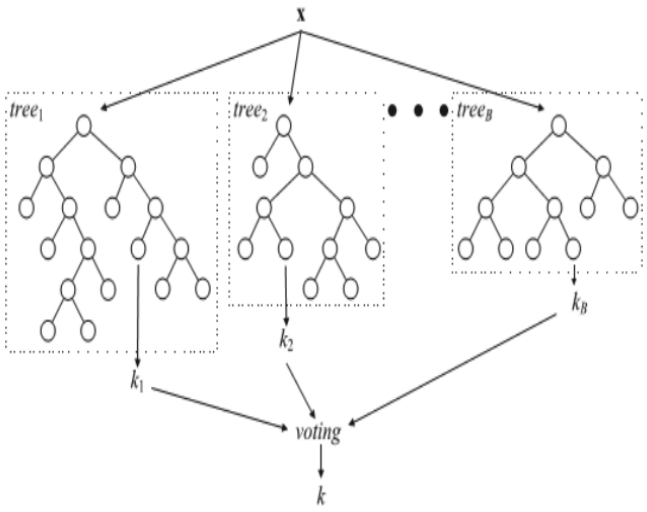

Fig. 5. Architecture of Random Forest.

E. Neural Network

The algorithm neural network is an artificial intelligence method whose concept is to apply a neural network system in the human body where nodes are connected to each other, architecture neural network as shown in Fig. 6.

The number of input layers before training is 11 input layers, after preprocessing the input layer to 17 input layers, in addition to determining the hidden layer, genetic algorithms on the neural network is used [24]. The GA-NN [25] algorithm process for this forecasting process is as follows:

- This forecasting is as follows:
- Initialization count = 0, fitness = 0, number of cycles
- Early population generation. Individual chromosomes are formulated as successive gene sequences, each encoding the input.
- Suitable network design
- Assign weights
- Conduct training with backpropagation Looks for cumulative errors and fitness values. Then evaluated based on the value of fitness.
- If the previous fitness <current fitness value, save the current value
- Count = count +1
- Selection: Two mains are selected using a wheel roulette mechanism
- Genetic Operations: crossover, mutation, and reproduction to produce new feature sets
- If (number of cycles <= count) return to number four
- Network training with selected features
- Study performance with test data.

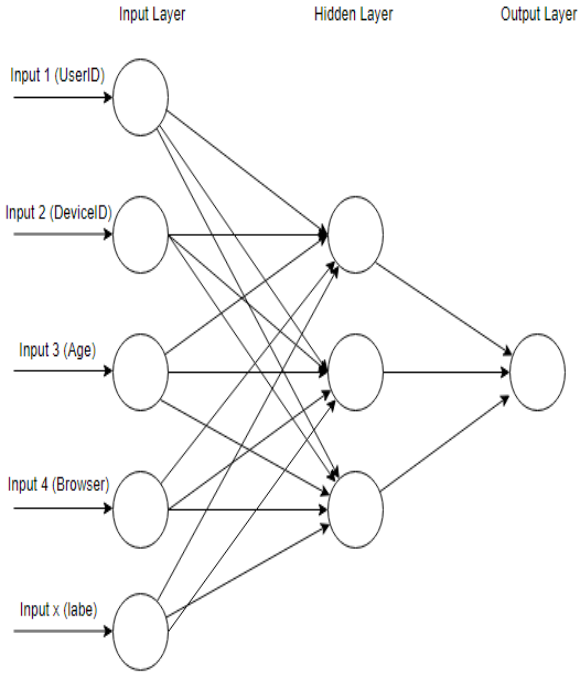

Fig. 6. Architecture of Neural Network.

F. Confusion Matrix

Confusion matrix is a method that can be used to evaluate classification performance. Table I shows a dataset with only two types of classes [26].

True Positive (TP) and True Negative (TN) are the number of positive and negative classes that are classified correctly, False Positive (FP) and False Negative (FN) is the number of positive and negative classes that are not classified correctly. Based on the confusion matrix, performance criteria such as Accuracy, Precision, Recall, F-Measure, G-Mean can be determined.

Accuracy is the most common criteria for measuring classification performance, but if working in an imbalanced class, this criterion is not appropriate because the minority class will have a small contribution to the accuracy criteria. The recommended evaluation criteria are recall, precision F-1 Score and G-Mean. F-1 Score is used to measure the classification of minority classes in unbalanced classes, and the G-mean index is used to measure overall performance (overall classification performance).

In this study, classification performance using Recall, Precision, F-1 Score and G-Mean:

$$\text{Accuracy} = \frac{\text{TP+ TN}}{\text{TP+T N+FN+FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TN+FP}} \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (4)$$

$$\text{G} - \text{Mean} = \sqrt{\text{TP} - \text{TN}} \quad (5)$$

$$\text{F1} - \text{Score} = \frac{\text{2 x Precision x Recall}}{\text{Precision + Recall}} \quad (6)$$

TABLE. I. CONFUSION MATRIX

Class	Predictive Positive	Predictive Negative
Actual Positive	TP	TN
Actual Negative	FP	FN

IV. RESULTS AND DISCUSSION

A. Dataset

This study uses an e-commerce fraud dataset sourced from Kaggle. The dataset consists of 151,112 records, a dataset classified as fraud is 14,151 records, and the ratio of fraud data is 0.093. SMOTE (Synthetic Minority Oversampling Technique) [27] minimizes class imbalance in the fraud transaction dataset by generating synthesis data, so that the total data consists of 151,112 records, dataset classified as fraud is 14,151 records, fraud data ratio is 0.093, as shown in Fig. 7.

After oversampling at the picture Fig. 8

The SMOTE (Synthetic Minority Oversampling Technique) process makes the synthesis data so that the data becomes balance.

B. Decision Trees

The experimental process using the decision tree model is done by preparing data that has been done by the preprocessing process. After preprocessing, the data will be carried out by oversampling the classification using the decision tree will be done using the oversampling data, and also the decision tree will be done by using the data that has not been oversampled. The results of these two experiments will show the results of the classification using a comparison of decision trees and the SMOTE (Synthetic Minority Oversampling Technique) oversampling process.

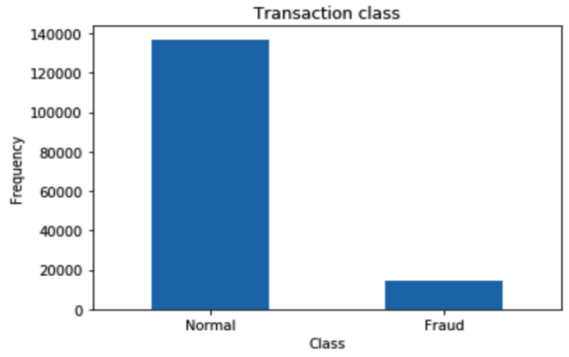

Fig. 7. Ratio Fraud.

Fig. 8. Ratio Fraud after over Sampling.

Decision tree without SMOTE produce Accuracy is 91%, recall is 59.8%, Precision is 54.1%, F1-Score is 56.8%, G-Mean is 75.2%. Table II shows result from confusion matrix decision tree without SMOTE.

Decision tree with SMOTE produce Accuracy is 91%, recallis 60.4%, Precisionis 91.6%, F1-Score is 91.2%, G-Mean is 75.3%. Table III shows result from confusion matrix decision tree with SMOTE.

C. Naïve Bayes

The process of testing using the Naïve Bayes model is done by preparing data that has already been done in the preprocessing process. After preprocessing, the data will be carried out oversampling using Naïve Bayes classification will be done using data that has been oversampling, and also Naïve Bayes will be done using data that is not oversampling. The results of these two experiments will show the results of the classification using the comparison of Naïve Bayes and the SMOTE (Synthetic Minority Oversampling Technique) oversampling process.

Naïve Bayes without SMOTE produce Accuracy is 95%, recall is 54.1%, Precision is 91.1%, F1-Score is 67.9%, G-Mean is 73.3%. Table IV shows result from confusion matrix naïve Bayes without SMOTE.

Naïve Bayes with SMOTE produce Accuracy is 95%, recall is 54.2%, Precision is 94.9%, F1-Score is 94.5%, G-Mean is 73.4%. Table V shows result from confusion matrix Naïve Bayes with SMOTE.

TABLE. II. CONFUSION MATRIX DECISION TREE WITHOUT SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	38782	38782
Actual Negative	1746	2595

TABLE. III. CONFUSION MATRIX DECISION TREE WITH SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	38651	2342
Actual Negative	1724	2617

TABLE. IV. CONFUSION MATRIX NAÏVE BAYES WITHOUT SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	40764	229
Actual Negative	1993	2348

TABLE. V. CONFUSION MATRIX NAÏVE BAYES WITH SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	40760	233
Actual Negative	1988	2353

D. Random Forest

The trial process using the Random Forest model is carried out by preparing data that has already been done by the preprocessing process. After preprocessing, the data will be carried out classification oversampling using Random Forest will be done using data that has been oversampled, and also Random Forest will be done using data that is not oversampling. The results of these two experiments will show the classification results using the Random Forest comparison and the SMOTE (Synthetic Minority Oversampling Technique) oversampling process.

Random forest without SMOTE produce Accuracy is 95%, recall is 55%, Precision is 95.5%, F1-Score is 69.8%, G-Mean is 74.0%. Table VI shows result from confusion matrix random forest without SMOTE.

Random Forest with SMOTE produce Accuracy is 95%, recall is 58.1%, Precision is 80.5%, F1-Score is 94.3%, G-Mean is 75.7%. Table VII shows result from confusion matrix random forest with SMOTE.

E. Neural Network

Research using the Neural Network model is done by preparing data that has already been done by the preprocessing process. After preprocessing, the data will be carried out classification oversampling using Neural Network will be done using data that has been oversampling, and also Random Forest will be done using data that is not oversampling. The results of these two experiments will show the results of classification using the Neural Network comparison and the SMOTE (Synthetic Minority Oversampling Technique) oversampling process.

Neural network without SMOTE produce Accuracy is 96%, recall is 54%, Precision is 97.1%, F1-Score is 97.1%, G-Mean is 73.5%. Table VIII shows result from confusion matrix neural network without SMOTE.

Neural network with SMOTE produce Accuracy is 85%, recall is 76.7%, Precision is 92.5%, F1-Score is 85.1%, G-Mean is 84.6%. Table IX shows result from confusion matrix neural network with SMOTE.

TABLE. VI. CONFUSION MATRIX RANDOM FOREST WITHOUT SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	40881	112
Actual Negative	1954	2387

TABLE. VII. CONFUSION MATRIX RANDOM FOREST WITH SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	40383	610
Actual Negative	1820	2521

TABLE. VIII. CONFUSION MATRIX NEURAL NETWORK WITHOUT SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	41113	24
Actual Negative	1932	2265

TABLE. IX. CONFUSION MATRIX NEURAL NETWORK WITH SMOTE

Class	Predictive Positive	Predictive Negative
Actual Positive	38566	2539
Actual Negative	9585	31487

Experiments using several algorithms produce accuracy values as shown in Fig. 9. The highest accuracy value in the neural network algorithm is 96%.

Experiments using several algorithms produce recall values as shown in Fig. 10, recall values increase using machine learning algorithms and also the Synthetic Minority Over Sampling Technique (SMOTE) compared only using the decision tree algorithm, random forest, Naïve Bayes, and neural networks only, the highest increase occurred in the neural network algorithm and the SMOTE (Synthetic Minority Over Sampling Technique).

Experiments using several algorithms produce precision values as shown in Fig. 11, the value decreases using machine learning algorithm and the Synthetic Minority Over Sampling Technique (SMOTE) compared only using the decision tree algorithm, random forest, Naïve Bayes, and neural networks, highest occurs in neural network algorithms and SMOTE (Synthetic Minority Over Sampling Technique).

Experiments using several algorithms produce F1-Score values as shown in Fig. 12, F1-Score values are increased by using machine learning algorithms and also Synthetic Minority Over Sampling Technique (SMOTE) compared only using algorithms. F1-Score is used to measure the classification of minority classes in unbalanced classes.

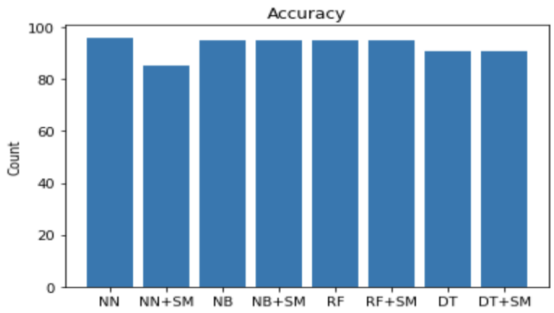

Fig. 9. Accuracy Result.

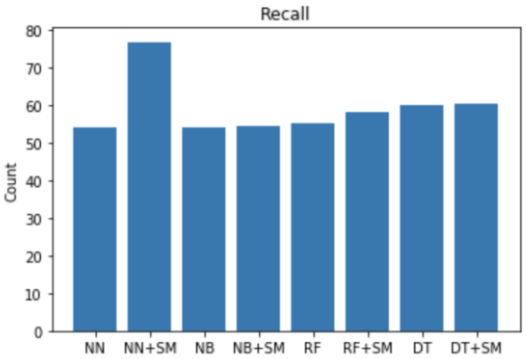

Fig. 10. Recall Result.

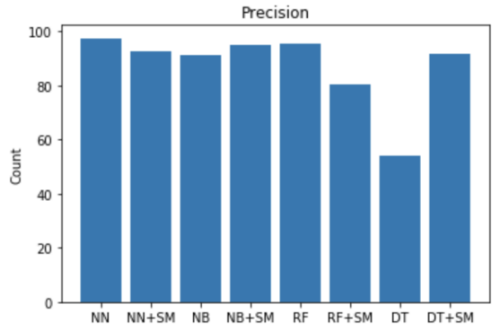

Fig. 11. Precision Result.

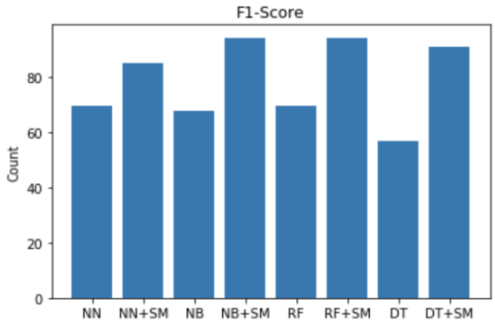

Fig. 12. F1-Score Result.

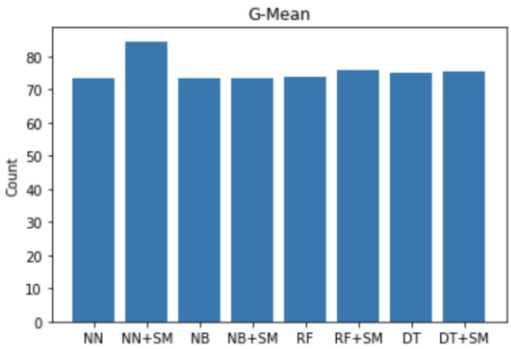

Fig. 13. G-Mean Result.

The G-mean value increased by using machine learning algorithm values as shown in Fig. 13, Synthetic Minority Over Sampling Technique (SMOTE) compared only using the G-mean algorithm used to measure overall performance (overall classification performance).

V. Conclusion

The e-commerce transaction fraud dataset is a database that has a class imbalance. This study applies the Synthetic Minority Over Sampling Technique (SMOTE) method to deal with class imbalances in the e-commerce transaction fraud dataset, the algorithm used is the decision tree, Naïve Bayes. random forest and neural network.

The results showed that the highest accuracy was 96% neural network, then random forest, and Naïve Bayes were 95%, for decision trees accuracy was 91%. Neural network has best accuracy because GA (genetic algorithms). Genetic algorithms can be used for improving ANN performance. Genetic algorithm can determine the number of hidden nodes and hidden layers, select relevant features, neural network. The SMOTE method in the experiment showed an increase in the value of recall, f1-score and also G -Mean, Neural network recall increased from 54% to 76.7%, Naïve Bayes recall increased from 41.2% to 41.3%, recall random forest from 55% to 58%, and recall decision tree from 59.8% to 60.4%. The value of f1-score also increased for all machine learning methods for neural networks increased from 69.8% to 85.1%, f1-score Naïve Bayes increased from 67.9% to 94.5%, f1-score random forest 69.8% to 94.3%, the f1-score for the decision tree also increased from 56.8% to 91.2%. By using SMOTE the value of G-Mean also increased for neural networks increased from 73.5% to 84.6%, G-Mean Naïve Bayes increased from 73.3% to 73.4%, G-Mean random forest 74% to 75 7%, the G-Mean for decision tree also increased from 75.2% to 75.3%.

Based on the results of the above experiment, it was concluded that the application of SMOTE on neural networks, random forests, decision trees, and Naïve Bayes was able to handle the imbalance of the e-commerce fraud dataset by producing higher G-Mean and F-1 scores compared to neural networks, random forest, decision tree, and Naïve Bayes. This proves that the SMOTE method is effective in increasing the performance of unbalanced data classification.

VI. Future Work

In Future studies, it is expected to be able to use other algorithms or deep learning for fraud detection in e-commerce and other future study to improve neural network accuracy when using the SMOTE (Synthetic Minority Over Sampling Technique) process.

References

[1] Asosiasi Penyelenggara Jasa Internet Indonesia, " Magazine APJI(Asosiasi Penyelenggara Jasa Internet Indonesia)" (2019): 23 April 2018.

[2] Asosiasi Penyelenggara Jasa Internet Indonesia, "Mengawali integritas era digital 2019 - Magazine APJI(Asosiasi Penyelenggara Jasa Internet Indonesia)" (2019).

[3] Laudon, Kenneth C., and Carol Guercio Traver. E-commerce: business, technology, society. 2016.

[4] statista.com. retail e-commerce revenue forecast from 2017 to 2023 (in billion U.S. dollars). (2018). Retrieved April 2018, from Indonesia: : https://www.statista.com/statistics/280925/e-commerce-revenue-forecast-in-indonesia/.

[5] Renjith, S. Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach. International Journal of Engineering Trends and Technology (2018).

[6] Roy, Abhimanyu, et al. "Deep learning detecting fraud in credit card transactions." 2018 Systems and Information Engineering Design Symposium (SIEDS). IEEE, 2018.

[7] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109-121.

[8] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109-121.

[9] Pumsirirat, Apapan, and Liu Yan. "Credit card fraud detection using deep learning based on auto-encoder and restricted boltzmann machine." International Journal of advanced computer science and applications 9.1 (2018): 18-25.

[10] Srivastava, Abhinav, et al. "Credit card fraud detection using hidden Markov model." IEEE Transactions on dependable and secure computing 5.1 (2008): 37-48.

[11] Lakshmi, S. V. S. S., and S. D. Kavilla. "Machine Learning For Credit Card Fraud Detection System." International Journal of Applied Engineering Research 13.24 (2018): 16819-16824.

[12] Aljarah, Ibrahim, Hossam Faris, and Seyedali Mirjalili. "Optimizing connection weights in neural networks using the whale optimization algorithm." Soft Computing 22.1 (2018): 1-15.

[13] Bouktif, Salah, et al. "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches." Energies 11.7 (2018): 1636.

[14] Xuan, Shiyang, Guanjun Liu, and Zhenchuan Li. "Refined weighted random forest and its application to credit card fraud detection." International Conference on Computational Social Networks. Springer, Cham, 2018.

[15] Hong, Haoyuan, et al. "Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China)." Catena 163 (2018): 399-413.

[16] Zhao, Jie, et al. "Extracting and reasoning about implicit behavioral evidences for detecting fraudulent online transactions in e-Commerce." Decision support systems 86 (2016): 109-121.

[17] Sharma, Shiven, et al. "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance." 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018.

[18] Kim, Jaekwon, Youngshin Han, and Jongsik Lee. "Data imbalance problem solving for smote based oversampling: Study on fault detection prediction model in semiconductor manufacturing process." Advanced Science and Technology Letters 133 (2016): 79-84.

[19] Sadaghiyanfam, Safa, and Mehmet Kuntalp. "Comparing the Performances of PCA (Principle Component Analysis) and LDA (Linear Discriminant Analysis) Transformations on PAF (Paroxysmal Atrial Fibrillation) Patient Detection." Proceedings of the 2018 3rd International Conference on Biomedical Imaging, Signal Processing. ACM, 2018.

[20] Harrison, Paula A., et al. "Selecting methods for ecosystem service assessment: A decision tree approach." Ecosystem services 29 (2018): 481-498.

[21] Randhawa, Kuldeep, et al. "Credit card fraud detection using AdaBoost and majority voting." IEEE access 6 (2018): 14277-14284.

[22] Lakshmi, S. V. S. S., and S. D. Kavilla. "Machine Learning For Credit Card Fraud Detection System." International Journal of Applied Engineering Research 13.24 (2018): 16819-16824.

[23] Li, Tong, et al. "Differentially private Naïve Bayes learning over multiple data sources." Information Sciences 444 (2018): 89-104.

[24] Suganuma, Masanori, Shinichi Shirakawa, and Tomoharu Nagao. "A genetic programming approach to designing convolutional neural network architectures." Proceedings of the Genetic and Evolutionary Computation Conference. ACM, 2017.

[25] Ruehle, Fabian. "Evolving neural networks with genetic algorithms to study the string landscape." Journal of High Energy Physics 2017.8 (2017): 38.

[26] Ting, Kai Ming. "Confusion matrix." Encyclopedia of Machine Learning and Data Mining (2017): 260-260.

[27] Siringoringo, Rimbun. "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor." Journal Information System Development (ISD) 3.1 (2018).