International Research Journal of Advanced Engineering and Science
ISSN (Online): 2455-9024

Analysis of Random Forest Algorithm on Customer Churn Prediction to Handle Imbalanced Data

Hafiz Ma'ruf[1], Rodiah[2]
[1]Master of Management Information System Department, Business Information System, Gunadarma University, Indonesia
[2]Informatic Department, Gunadarma University, Indonesia
Email Address: [1]hafizmrf3 @ gmail.com, [2]rodiah @ staff.gunadarma.ac.id

Abstract— *One way to understand the customer's interest in a product is to learn the customer buying behavior pattern when making transactions of a product. This customer buying behavior pattern is considered crucial for the company to comprehend the level of customer satisfaction and comfort of the products or services provided. Companies that implement a data-driven mindset will be better in understanding the characteristics of customers that might switch to the products of other companies or competitors (customer churn). In this study, the implementation of Random Forest algorithm is conducted to predict customer churn on e-commerce retail data. The research stage in this study is focused on Data Understanding, Data Preparation, Modeling, and Evaluation. This study also conducted SMOTE techniques to handle imbalanced data, so as to obtain higher accuracy than before. The predictive results are then evaluated with confusion matrix. The first scenario model uses a combination of numerical and categorical variables managed to become the best performing model because it achieved 95.41% accuracy in the test data. The second scenario model that uses only categorical variables is only able to produce accuracy of 89.43% in the test data. The results of this study are expected to learn customer patterns in order to increase customer loyalty to a product.*

Keywords— *Customer Churn; E-commerce; Imbalanced Data, Random Forest, SMOTE.*

I. Introduction

Trade transactions that were originally done by meeting directly between sellers and buyers, are now starting to change with the use of e-commerce. The dissemination of information about a product can be done faster and has a very wide scope, so this began to shift the pattern of consumption and has even now become part of the lifestyle of society. Businesses that use the internet to receive orders, or conduct sales of goods and or services in 2020 amounted to 90.18% [1]. The e-commerce industry in Indonesia is currently doing a lot of innovations followed by increasingly intense competition. This is due to the dynamic change of customers in choosing one or more of the many existing service providers. Customer churn is a situation where the customer's contribution to the company's profit decreases [2]. Churn rate is considered an important metrics for the company because it affects the company's revenue. One of the reasons why customer churn happens is because the customer is no longer interested in the product or service and has found an alternative solution from competitors [3]. In this market competition is seen from the experience that each year about 30-35% churn rate and requires 5-10 times the effort and cost to add new customers rather than maintaining existing ones [4].

In terms of retaining customers, companies need a way to find out what factors or customer characteristics will churn. One way to predict customer churn is to implement the use of machine learning algorithms. With machine learning, companies will find it easy to find the right solution based on existing historical data. Research [5] predicts customer churn based on datasets obtained using online data crawling techniques from one of the E-commerce companies. The dataset consists of 5 variables (Price, FirstTimeDiff, TimeDiff, Frequency, Score, and Label). The model is built using the Neural Networks algorithm with an accuracy of 82.64%. Research [6] predicts customer churn based on statistical data obtained at one of the telecommunications companies. The dataset consists of 17 variables that are generally related to the daily use of customer service, international calls, and customer service calls. The modeling was carried out using 10 algorithms, where the model with the Random Forest algorithm managed to achieve the highest accuracy of 96%. Research [7] predicts customer churn based on a dataset that is divided into 5 types; Customer Data, Tower and Complaint Database, Network Log Data, Call Detail Records (CDR), and Mobile IMEI Information. The model with XGBoost algorithm succeeded in producing the best accuracy with AUC of 93.3%. Research [8] predicts customer churn based on a dataset obtained from the database of the American Telcom company, Orange. The dataset consists of 20 of which only 11 variables are used for the modeling stage. Of the 2 algorithms applied, the best model is the Logistic Regression algorithm which produces an accuracy of 85.24%. Research [9] predicts customer churn based on a dataset obtained using the online data crawling technique of one bank which consists of 57 variables that can generally be grouped into demographics, transactions, and balances. Of the 5 algorithms applied to modeling, the SVM algorithm managed to become the best algorithm with an accuracy of 92.65%.

In this study, the implementation of the Random Forest algorithm to predict and classify customer churn was carried out on a dataset obtained from the public dataset site Kaggle.com. The results of the classification will be tested with the Confusion Matrix to see how much accuracy is in classifying customer churn. The accuracy of the research was then improved by applying the Synthetic Minority Over-sampling Technique (SMOTE) technique.

II. Methods

In this study, Random Forest algorithm was implemented to predict customer churn. The dataset used in this study belongs to an online retail company (e-commerce) with B2C model which is then published to the public on the public dataset site Kaggle.com. The initial stage is done by initializing research variables and standardization of data to prevent over-fitting of models. The next process is done splitting data into training data and test data to then be modeled with Random Forest algorithm. The final result of this study is in the form of accuracy of customer churn predictions. An overview of this study method can be seen in figure 1.

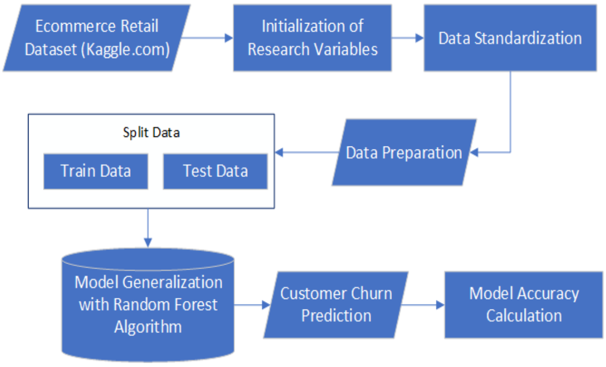

Ecommerce Retail Dataset (Kaggle.com)
Initialization of Research Variables
Data Standardization
Split Data
Train Data
Test Data
Data Preparation
Model Generalization with Random Forest Algorithm
Customer Churn Prediction
Model Accuracy Calculation

Fig. 1. Research Stages.

A. *Data Understanding*

The dataset used in this study is the first version published by Ankit Verma and last updated January 26, 2021. The dataset contains retail e-commerce customer data in excel workbook (.xlsx) type format consisting of 5630 records and 20 variables. Dataset is divided into 2 variable types, where 11 variables with categorical type and 9 variables with numerical type.

TABLE I. Variable Descriptions.

Variable	Type	Description
CustomerID	Categorical	ID of customer
Churn	Categorical	Does customer churn (potentially switching to another e-commerce) (1 means Yes, 0 means No)
Tenure	Numerical	Customer subscription duration (Month)
PreferredLogin Device	Categorical	Device used to Login the Application
CityTier	Categorical	Level of customer proximity to the city
Warehouse ToHome	Numerical	Distance between warehouse to customer's house (KM)
PreferredPayment Mode	Categorical	Types of payments commonly used by customers
Gender	Categorical	Customer’s gender (Male, Female)
HourSpend OnApp	Categorical	Duration of customer using the app (hours)
NumberOfDevice Registered	Numerical	Number of devices registered with customer’s account
PreferredOrderCat	Categorical	Last month customer’s order category during transaction
SatisfactionScore	Categorical	Customer satisfaction rating on the service provided (1.0 – 5.0)
MaritalStatus	Categorical	Is the customer married (1 means Yes, 0 means No)
NumberOf Address	Numerical	Number of addresses added by a customer
Complain	Categorical	Is there a complaint from the customer (1 means Yes, 0 means No)
OrderAmount HikeFrom LastYear	Numerical	Increase in Number of Orders by customers from the previous year
CouponUsed	Numerical	Number of coupons that have been used by customers in the last month
OrderCount	Numerical	Number of orders made by customers in the last month of transaction
DaySinceLast Order	Numerical	Number of days since the last customer order was made
CashbackAmount	Numerical	The amount of cashback earned by customers in the last month of the transaction

At this stage, a more detailed understanding of the dataset is used. Starting from initialization of variables to data exploration to know customer characteristics, including univariate and bivariate analysis. Univariate analysis in this study is conducted by analyzing the number of customer and churn rate based on categorical variables and numerical variables. Bivariate analysis is done by looking for correlation values between variables.

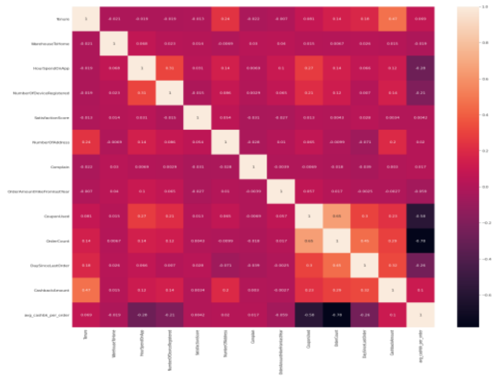


Fig. 2. Correlation Heatmap.

Based on the results of heatmap visualization in figure 2, 3 of the highest correlation values can be taken:

1. Correlation between OrderCount and CouponUsed is 0.65
2. Correlation between CashbackAmount and Tenure is 0.47
3. Correlation between Ordercount and DaySinceLastOrder is 0.45.

The correlation value between variables will always be in the range of values between -1 and +1 [10]. The highest correlation value is in the OrderCount and CouponUsed variables which have a correlation value of 0.65, which means it is close to 0.7 to state the moderate positive relationship between the two variables [11]. It also concludes that customers have a habit of ordering accompanied by the use of coupons on e-commerce applications.

Then on the variable CashbackAmount and Tenure has a correlation value of 0.47, which means that customers who often get cashback will have an impact on the length or loyalty of the customer in subscribing to the company's e-commerce service. Furthermore, the OrderCount and DaySinceLastOrder variables have a correlation value of 0.45 which is enough to provide information that the high number of orders made by customers is usually a pattern based on the range of days after the order is placed on the e-commerce application.

B. *Data Preparation*

Data preparation in this study was conducted to ensure the data to be processed at the modeling stage is good quality, so as to minimize overfitting or underfitting models in predicting customers who churn or not. This step starts with variable selection, null values handling, outliers handling, label encoding, and ends with data standardization. These stages use the help of pandas library and sklearn.preprocessing library from scikit-learn library.

The CustomerID and MaritalStatus variables in this study were considered to have no impact on predictive results. Therefore both variables need to be removed so that the quality of the model becomes better when predicting. Checking the number of null values in each variable is done to find out how much data contains null values. The steps taken to handle null values in this study by replacing null values with the median values of each variable. The median value is considered as a solution to replace null values because of the data distribution tends to be skewed to the right (positive skewness) [12].

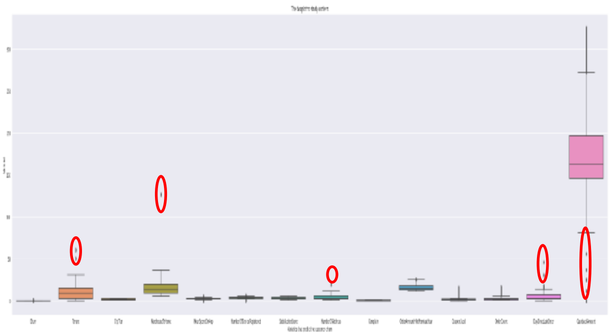


Fig. 3. Outlier Visualization.

Based on the outlier visualization in figure 3, it can be seen through the image circled in red that there are outliers on some numerical variables. Therefore it needs to be addressed by removing the outlier and the result of outlier handling can be done with the following pseudocode:

```
remove_outlier
Var Q1,Q2 = percentile(25,75)
Algorithm
IQR = Q3-Q1
lr = Q1 – (1.5 * IQR)
ur = Q3 + (1.5 * IQR)
return lr, ur
```

The data in this study have different units of measurement so that the original data must be transformed before modeling can be done. The purpose of standardization is to equalize the unit so that the value no longer depends on the unit of measurement but rather becomes the standard value. This standardization is done based on the formula of equation 1 [13].

$$x_{new} = \frac{x - \mu}{\sigma} \qquad (1)$$

C. *Split Data*

The model test design is done with 2 modeling scenarios. The model with the first scenario, all variables both numerical and categorical are processed into the model, then done splitting data where 80% for the training data and 20% for the test data, then tested and compared the accuracy results before and after applied SMOTE technique. While in the model with the second scenario, only categorical variables are processed into the model, then done splitting data where 80% for the training data and 20% for the test data, then tested and compared the accuracy results before and after the SMOTE technique is applied. The test scenario was conducted to determine how much influence the variables used on the data would be trained and tested on the accuracy obtained by the model.

D. *Model Generalization with Random Forest*

The process of the Random Forest algorithm is the same as used by the Classification And Regression Tree (CART) [14] (Han, et al., 2012). CART uses gini index to measure the selection of attributes that will be used to separate each class using attributes from e-commerce dataset. The attribute with the highest gini index will be used to separate n nodes. Gini index can be calculated using the equation 2 [15]:

$$Gini\ Index = 1 - \sum_{i=0}^{c} P_i^2 \qquad (2)$$

Where:
c = class
p_i = probability of class i

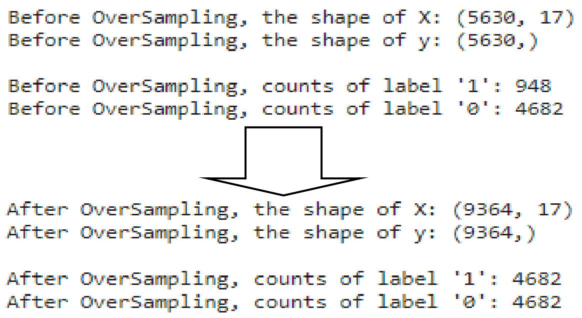
The model is built with several parameters. The number of trees in the forest as many as 100. The criterion of a split quality is ‘gini’ for the gini impurity. The maximum depth of the tree is 10. If it does not exist, then the node is expanded until all the leaf is pure or until all the leaf is less than the sample min_samples_split (the minimum number of samples required to divide the internal node/decision node).

E. *Model Evaluation*

In this study, evaluation was conducted using confusion matrix. Where the result is a specification table used to measure the accuracy performance of a model by comparing the results of classification conducted by the system with the actual classification results. From as many as 5630 customer data divided into 2 classes. Churn class as many as 948 customers, while the class does not churn as many as 4682 customers. This problem greatly affects the predictive performance of classification algorithm models because models tend to predict with a high degree of accuracy in larger classes the amount of data (major class) [16]. Based on the data level approach, there are various data resampling and duplication techniques to balance minor class data on trained data [17]. One technique is oversampling by balancing classes by randomly duplicating minor classes. The drawback of oversampling is that it will experience overfitting due to the exact same minor class duplication [18].

The Synthetic Minority Oversampling Technique (SMOTE) was then conducted in this study to address and reduce overfitting [19] which is the weakness of oversampling techniques, namely by utilizing the nearest neighbor (k-

Nearest Neighbor) of the desired number of oversampling. The research was then resumed at the stage of data preparation and data oversampling with SMOTE technique. This step is done with the help of the SMOTE() function of the imblearn.over_sampling package.



```
Before OverSampling, the shape of X: (5630, 17)
Before OverSampling, the shape of y: (5630,)

Before OverSampling, counts of label '1': 948
Before OverSampling, counts of label '0': 4682

After OverSampling, the shape of X: (9364, 17)
After OverSampling, the shape of y: (9364,)

After OverSampling, counts of label '1': 4682
After OverSampling, counts of label '0': 4682
```

Fig. 4. SMOTE Implementation.

Based on figure 4, the imbalanced data in Churn variable is then applied SMOTE by oversampling the amount of data on the label '1' (Churn) and equating to the amount of data in label '0' (Not Churn), so that the number of rows of data on the label '1' and '0' is balanced and each class contains 4682 data. The dataset used in the modeling stage becomes 9364 rows of data.

III. RESULTS AND DISCUSSION

Manual calculation of Random Forest algorithm by calculating gini index value successfully. For example, a calculation on a SatisfactionScore variable with data in table II.

TABLE III. SatisfactionScore

Satisfaction Score	Yes Churn	Not Churn	Total
1	134	1030	1164
2	74	512	586
3	292	1406	1698
4	184	890	1074
5	264	844	1108
Total	948	4682	5630

Based on table II,the gini index value of SatisfactionScore variable can be calculated by equation 2 as follows:

$$\text{Gini(Score1)} = 1 – (((134/1164)^2) + ((1030/1164)^2))$$
$$\text{Gini(Score1)} = 0.204$$
$$\text{Gini(Score2)} = 1– (((74/586)^2) + ((512/586)^2))$$
$$\text{Gini(Score2)} = 0.221$$
$$\text{Gini(Score3)} = 1 – (((292/1698)^2) + ((1406/1698)^2))$$
$$\text{Gini(Score3)} = 0.285$$
$$\text{Gini(Score4)} = 1 – (((184/1074)^2) + ((890/1074)^2))$$
$$\text{Gini(Score4)} = 0.284$$
$$\text{Gini(Score5)} = 1 – (((264/1108)^2) + ((844/1108)^2))$$
$$\text{Gini(Score5)} = 0.363$$
$$\begin{aligned}\text{Gini(SatisfactionScore)} = {} & ((1164/5630) * \text{Gini(Score1)}) + \\ & ((586/5630) * \text{Gini(Score2)}) + \\ & ((1698/5630) * \text{Gini(Score3)} + \\ & ((1074/5630) * \text{Gini(Score4)}) + \\ & ((1108/5630) * \text{Gini(Score5)})\end{aligned}$$
$$\text{Gini(SatisfactionScore)} = 0.277$$

Calculations on other variables use the same equation, but are not explained in more detail.

Modeling was successfully implemented using the Random Forest algorithm with 2 scenarios. The model with the first scenario, all variables both numerical and categorical are processed into the model, then done splitting data where 80% for the training data and 20% for the test data, then tested and compared the accuracy results before and after applied SMOTE technique. While in the model with the second scenario, only categorical variables are processed into the model, then done splitting data where 80% for the training data and 20% for the test data, then tested and compared the accuracy results before and after the SMOTE technique is applied. Accuracy results are obtained with the help of built-in functions confusion_matrix() and classification_report() of package sklearn.metrics..

In tables III, IV, V, and VI, it can be seen that the yellow cell represents the True Positive value and the True Negative value of the classification result applied by the Random Forest model to the test data.

TABLE IIIII. Scenario 1 Before SMOTE

Confusion Matrix on Test Data	**Predicted (Not Churn)**	**Predicted (Churn)**
Actual (Not Churn)	931	7
Actual (Churn)	63	125

From 1126 test data in scenario 1, the model can predict with 93.78% accuracy.

TABLE IVV. Scenario 1 After SMOTE

Confusion Matrix on Test Data	**Predicted (Not Churn)**	**Predicted (Churn)**
Actual (Not Churn)	901	46
Actual (Churn)	40	886

From 1873 test data in scenario 1, the model can predict with 95.41% accuracy after SMOTE technique applied in dataset.

TABLE V. Scenario 2 Before SMOTE

Confusion Matrix on Test Data	**Predicted (Not Churn)**	**Predicted (Churn)**
Actual (Not Churn)	922	16
Actual (Churn)	103	85

From 1126 test data in scenario 2, the model can predict with 89.43% accuracy.

TABLE VI. Scenario 2 After SMOTE

Confusion Matrix on Test Data	**Predicted (Not Churn)**	**Predicted (Churn)**
Actual (Not Churn)	830	117
Actual (Churn)	102	824

From 1873 test data in scenario 2, the model can predict with 88.31% accuracy after SMOTE technique applied in dataset.

To make it easier to compare the results of modeling before and after the data is applied SMOTE technique, the evaluation conducted in this study is by presenting the model performance data in the form of bar charts in figure 5.

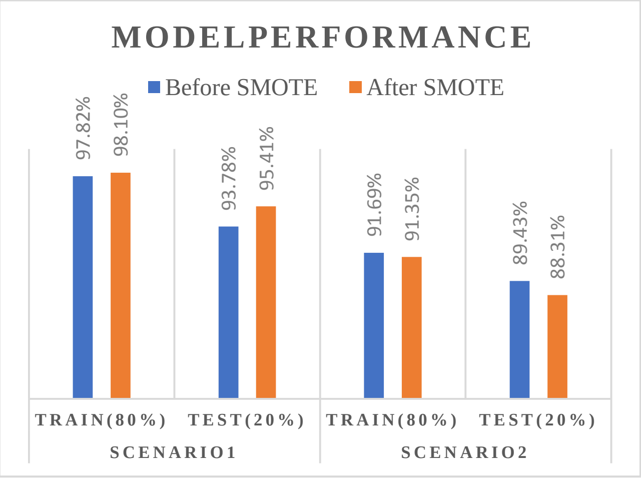


Fig. 5. Model Performance

Data visualization shown in figure 4.4 is the final result of a series of processes that have been performed using the Random Forest algorithm with the first scenario and the second scenario. It can be seen in scenario 1, that the model that processes all variables, both numerical and categorical managed to be the best model capable of achieving 98.10% accuracy on the training data and 95.41% in the test data. In scenario 2, models that process categorical variables alone are only able to achieve 91.69% accuracy in the training data and 89.43% in the test data. It is also seen that the effect of this SMOTE technique if applied only to categorical variables and the accuracy results obtained tend to decrease by about 2-3%. Then it can be concluded that the model with the first scenario is considered as the best model in this study to get equally good accuracy value on the training data and test data. So this model is worth using to predict customer churn.

IV. Conclusions and Suggestions

Based on the results and discussion, several conclusions can be drawn. The prediction customer churn model using Random Forest algorithm was successfully done with the help of numpy, pandas, scikit-learn, and matplotlib libraries. The accuracy calculation on the training data and test data from all three models were successfully performed using confusion matrix. Improved model accuracy was successfully performed by applying the SMOTE technique to the dataset. The first scenario became the best algorithm in this study because it managed to achieve 95.41% accuracy in the test data. So it deserves to be a model that can be deployed into a real-time application.

Further development can be made to improve research by applying data preparation that is much more detailed neural networks algorithms or in more detail with deep neural networks to be able to produce higher accuracy. Then the development of further research can include the stage of deployment model to know in real-time application whether the model that has been built today has answered the needs of business in predicting customer churn.

Acknowledgment

Thanks to SarMag Program and Master of Management Information System Department, Gunadarma University, Depok, Indonesia.

References

[1] Kusumatrisna, A. L., Rozama, N. A., Syakilah, A., Wulandari, V. C., Untari, R., Sutarsih, T, “Statistik E-Commerce 2020,” Jakarta: Badan Pusat Statistik. 2020.

[2] Xiaobing, Y., Jie, C., and Zaiwu, G. “The Review of Customer Churn Issue,” *Computer Integrated Manufacturing System*, vol. 10, pp. 2253-2263, 2012.

[3] Fadiyah, S. "Pengertian Churn Rate dan Penjelasan Lengkapnya," Available at: https://www.hashmicro.com/id/blog/pengertian-churn-rate-dan-penjelasan-lengkapnya/ [Accessed 29 April 2021], 2020.

[4] Hanifa, T. T., A. and Al-Faraby, S. “Analisis Churn Prediction pada Data Pelanggan PT. Telekomunikasi dengan Logistic Regression dan Underbagging,” *Telkom University*, vol. 4, issue 2, pp. 3210-3225, 2017.

[5] Xia, G. and He, Q. “The Research of Online Shopping Customer Churn Prediction Based Integrated Learning, ” *Atlantis Press*, vol. 149, pp. 756-764, 2018. doi: 10.2991/mecae-18.2018.133.

[6] Sabbeh, S. F. “Machine-Learning Techniques for Customer Retention: A Comparative Study,” *International Journal of Advanced Computer Science and Applications*, vol. 9 issue 2, pp. 273-281, 2018. doi: 10.14569/IJACSA.2018.090238.

[7] Ahmad, A. K., Jafar, A. and Aljoumaa, K. “Customer Churn Prediction in Telecom Using Maching Learning in Big Data Platform,” *Journal of Big Data*, vol. 6 issue 28, pp. 1-24, 2019. doi: 10.1186/s40537-019-0191-6.

[8] Jain, H., Khunteta, A. and Srivastava, S. “Churn Prediction in Telecommunication using Logistic Regression and Logit Boost,” *International Conference on Computational Intelligence and Data Science*, Jaipur, India, pp. 101-112, 2019. doi: 10.1016/j.procs.2020.03.187.

[9] Karvana, K. G. M., Yazid, S., Syalim, A. and Mursanto, P. “Customer Churn Analysis and Prediction Using Data Mining Models in Banking Industry,” *International Workshop on Big Data and Information Security 2019*, Bali, Indonesia, pp. 33-38, 2019. doi: 10.1109/IWBIS.2019.8935884.

[10] Samuel, M. & Okey, L. E. “The Relevance and Significance of Correlation in Social Science Research,” *International Journal of Sociology and Anthropology Research*, vol. 1, issue 3, pp. 22-28, 2015.

[11] Ratner, B. “The Correlation Coefficient: Its Values Range Between +1/-1, or do they?,” *Journal of Targeting, Measurement, and Analysis for Marketing*. vol. 17, pp. 139-142, 2009. doi: 10.1057/jt.2009.5.

[12] Kim, H. Y. “Statistical Notes for Clinical Researchers: Assesing Normal Distribution (2) using Skewness and Kurtosis,” *Restor Dent Endod*. vol. 38 issue 1, pp. 52-54, 2013. doi: 10.5395/rde.2013.38.1.52.

[13] Alexandropoulos, S., Kotsiantis, S., and Vrahatis, M. “Data preprocessing in predictive data mining,” *The Knowledge Engineering Review*, vol. 34, ed. 1, 2019. doi:10.1017/S026988891800036X.

[14] Han, J., Kamber, M. and Pei, J. “Data Mining: Concepts and Techniques”, 3 rd ed. Burlington: Morgan Kaufmann, 2012.

[15] Sharda, R., Delen, D., and Turban, E. “Business Intelligence and Analytics: Systems for Decision Support,” 10th ed. London: Pearson, 2013.

[16] Sanguanmak , Y. and Hanskunatai, A. “Auto-tuning of parameters in hybrid sampling method for class imbalance problem”. 2016 International Computer Science and Engineering Conference, pp. 1-5, 2016. doi: 10.1109/ICSEC.2016.7859941.

[17] Zhang, D., Liu, W., Gong, X. and Jin, H. “A Novel Improved SMOTE Resampling Algorithm Based on Fractal”. *Journal of Computer information Systems*, vol. 7 issue 6, pp. 2204-2211, 2011.

[18] Yap, B. W., Rani, K. A., Rahman, H. A. and Fong, S. “An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets,” *Proceedings of the First International Conference on Advanced Data and Information Engineering*, Singapore, Springer, vol. 285, pp. 13-22, 2014. doi: 10.1007/978-981-4585-18-7_2

[19] Shen, L., Lin, Z. and Huang, Q., “Relay Backpropagation for Effective Learning of Deep Convolutional Neural Networks”. Cham, Springer, 2016.