Evaluation of Decision Tree, K-NN, Naive Bayes and SVM with MWMOTE on UCI Dataset

Meida Cahyo Untoro[1*], Mugi Praseptiawan[1], Mastuti Widianingsih[2], Ilham Firman Ashari[1], Aidil Afriansyah[1], and Oktafianto[1]

[1]Department of Informatics engineering, Institut Teknologi Sumatera, Indonesia
[2]Department of Biology, Universitas Gadjah Mada, Indonesia

*cahyo.untoro@if.itera.ac.id

Abstract. Imbalanced data causes misclassification because the majority of the dominant data is in the minority data, which results in a decrease in the value of accuracy. UCI dataset is a public dataset that can be used as a dataset in machine learning. This study aims to evaluate the Decision Tree, K-NN, Naive Bayes, and Support Vector Machine classification methods on data imbalances in MWMOTE. MWMOTE is used in resolving Imbalanced cases through weighting and grouping. This goal is achieved by evaluating the Decision Tree, K-NN, Naive Bayes, and Support Vector Machine classification methods in MWMOTE to produce more representative synthetic data and increase the accuracy value. The results obtained from this study indicate that the Decision Tree has higher evaluations of recall, precision, F-measure, and accuracy compared to K-NN, Naive Bayes, and Support Vector Machine for data that are balanced with MWMOTE.

1. Introduction

Misclassification is a problem that often occurs in classifying Imbalanced data because classifiers are more inclined towards majority data so that low accuracy is obtained in minority data [1]. To handle imbalanced, some research manipulates data samples (synthetic data creation) and the use of algorithms [2]. Classification methods provide accuracy values for all data by eliminating minority classes and all data considered as the majority class. The dataset is assumed to have a balanced distribution, and minority classes will be noise or outliers [3][4]. Imbalanced data problems between minority and majority data, causing minority data accuracy to be low [5]. Imbalanced distribution results in classification events that are more inclined to the majority of data (negative) compared to the number of minority data (positive) [6].

State that the case of misclassified is caused by the imbalanced dataset [7]. Imbalanced cases can group data into 2, namely minority and majority data [2]. Also, imbalanced can lead to poor model making [8] as well as overfitting and decreasing classification accuracy [9]. Oversampling is one way to handle imbalanced problems by distributing balanced data by randomly replicating minority (synthetic data) data by iterating. Oversampling has disadvantages in making synthetic data with the appearance of overfitting because this mechanism makes synthetic data less precise. The Majority Weighted Minority Oversampling Technique (MWMOTE) can handle overfitting. Making synthetic data in MWMOTE has three stages, namely identification of minority class samples and majority classes on datasets, minority class weighting, and clustering The results of these proposals were able to reduce the degree of bias or noise and to produce synthetic data with better accuracy [2]. In this study, using Decision Tree, K-NN, Naive Bayes, and Support Vector Machine in classifying imbalanced data using MWMOTE in the UCI dataset, especially in pre-processing and testing phases.

2. Related Work

2.1. Decision Tree

Decision Tree or decision tree method is an algorithm of ID3 development that is used to predict data or facts large enough to become a decision tree by classifying or segmenting or increasing prediction [10]. To select an attribute as the root, based on the highest gain value (**1**) of the existing attributes. After getting the gain value, there is one more thing that needs to be done which is to calculate the value of entropy (2). Entropy is used to determine how informative an input attribute is to produce an output attribute [11].

$$Gain\ (F) = Entropy\ (D)\ - \sum_{b=1}^{a} \times\ Entropy \qquad (1)$$

D = Dataset
F = Total Dataset
a = Total of Feature D
Db = Total case to b on D

$$Entropy\ (D)\ = \sum_{b=1}^{a} -\ pb \times \log 2\ pb \qquad (2)$$

D = Dataset
a = Total of Feature D
pb = Probabilities D

2.2. K-Nearest Neighbors

K - Nearest Neighbors (K-NN) is one of the simple algorithms in the learning algorithm to predict a class in a dataset [12]. Classification of classes on K-NN based on the closest neighbours distance using Euclidean distance, City block distance, Cosine distance, Correlation, Hamming distance [13]. The distance between neighbours is an important part to optimize the K-NN algorithm, so the authors use Euclidean distance (**3**) [14]. The lack of the K-NN algorithm requires store or memory and the computational process is quite large [15].

$$Distance = \sqrt{(Lat_1 - Lat_2)^2 + (Long_1 - Long_2)^2} \qquad (3)$$

2.3. Naïve Bayes

Where to classify, data must be provided that have been defined for each attribute or class of criteria and classes [16]. To do the classification is calculated based on the probability value of each class for the variable (4).

$$p(H|\ E) = \frac{p(H) \times p(H|E)}{p(E)} \qquad (4)$$

p (H | E) = hypothesis probability value for evidence, p (H) = hypothesis probability value, p (E | H) = probability evidence value for the hypothesis, and p (E) = probability value of evidence.

2.4. Support Vector Machine

Support Vector Machine is a learning machine algorithm that works on the principle of Structural Risk Minimization (SRM) to find the best hyperplane (**Figure 1**) that separates two classes in the input space [10].

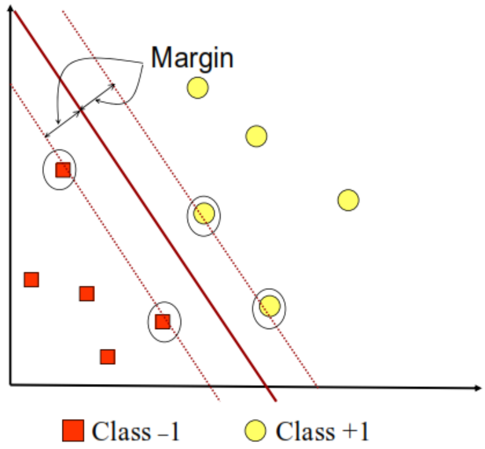

Figure 1. Hyperplane support vector machine

2.5. Imbalanced Data

Data that has Imbalanced ratio between one data and other data can be said to be imbalanced. Data mining means imbalanced by the amount of majority class data more than the minority class. Imbalanced problems occur in machine learning so that often results in misclassification has an impact on the value of the accuracy of class predictions decreases [17]. The decrease in accuracy in imbalances is due to the presence of noise or outliers in test datasets from minority classes [18]. One way to deal with imbalanced is by comparing classification methods with the addition of algorithms or modifying methods [19]. Imbalanced can be solved by adding synthetic data to the minority class with the method of oversampling and under-sampling. Imbalanced has quite high complexity and is differentiated into 3 cases (**Figure 2**) [20].

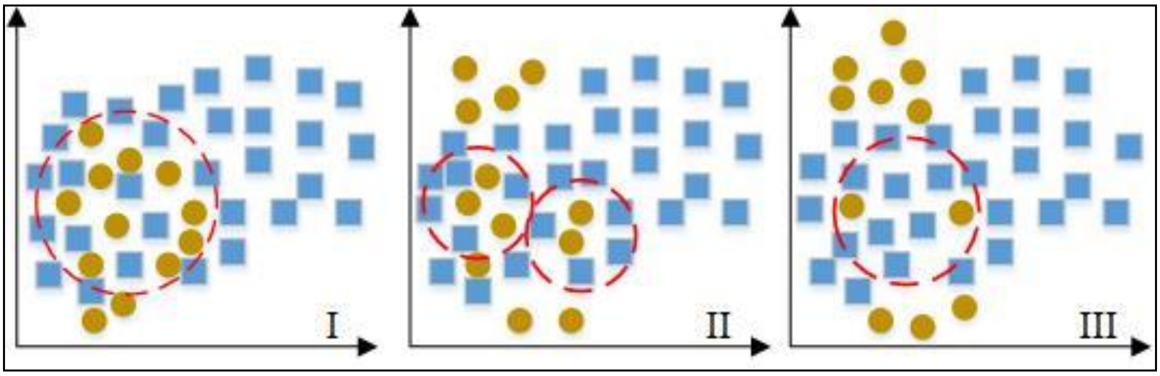

Figure 2. Imbalanced problem

3. Methodology

3.1. Dataset UCI

Imbalanced data set is a special case for classification problems where class distribution is not uniform among classes. Usually, they are organized by two classes: the majority (negative) and minority (positive) classes [21] (table 1).

Table 1. Dataset UCI

No	Dataset	Attribute	Examples	Imbalance Ratio
1	Abalone	8	731	0.94:0.06
2	Breast	10	106	0.66:0.34
3	E.coli	8	336	0.77:0.23
4	Robot	25	5456	0.78:0.22
5	Yeast	9	1484	0.79:0.21

3.2. Oversampling

The oversampling method for making synthetic data may have some inaccuracies in many scenarios. To overcome this problem, a new method of Majority Weighted Minority Oversampling Technique (MWMOTE) [2]. The purpose of MWMOTE is twofold, namely: To improve the process of sample selection and to improve the process of making synthetic samples. MWMOTE has three stages (**Figure 3**), namely: MWMOTE identifies minority data that is difficult to study. Minority data that is in the majority data, adjacent minority data (borderline) with the majority data and minority data that information is on the borderline. Second, each member of an informative minority sample is assigned a weighted sample selection weight (Sw).

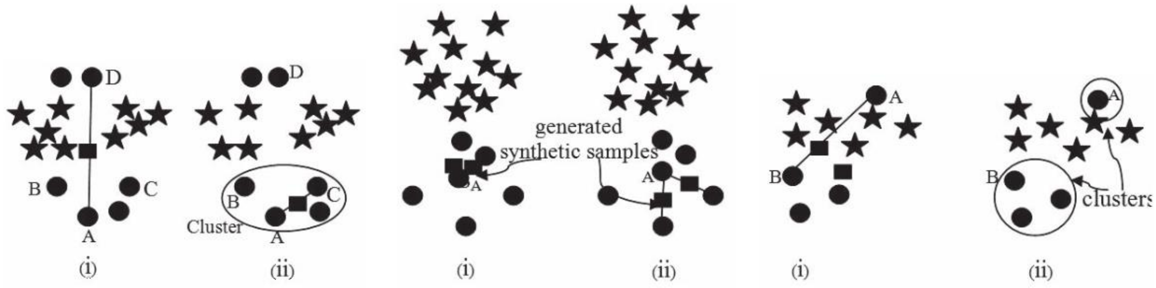

Figure 3 Oversampling MWMOTE [2]

3.3. Classification

Classification is the process of creating models using data testing and separating dataset categories by labelling each class [22]. The purpose of classification can be used as a prediction for future data trends [10]. The stages in classification consist of three parts [23], for the stage of model development, a model is created to solve the problem of classifying attributes or classes in a dataset. The model is built based on training data sets of problems faced and has good information. The stage of applying the built model is used to determine the attribute or class of testing data with the attribute/class not yet known. Evaluation is the stage of applying the previous model evaluated using measured parameters to determine whether the model is acceptable or not.

3.4. Evaluation

Performance evaluation uses precision, recall, F-Measure, and accuracy for each class, using True positive (TP), True negative (TN), False positive (FP), False Negative (FN).

$$Precision = \frac{TP}{(TP + FP)} \tag{5}$$

$$Recall = \frac{TP}{(TP + TN + FP + FN)} \tag{6}$$

$$F - Measure = 2 \times \frac{Precision.Recall}{(Precision + Recall)} \tag{7}$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{8}$$

Precision is the level of accuracy between the information requested by the user and the answers provided by the system. Recall is the level of success of the system in finding back information. F-measure is one of the evaluation calculations in the information retrieval that combines recall and precision. Accuracy is defined as the level of closeness between the predicted value and the actual value.

4. Result and Discussion

In this study, the evaluation of Decision Tree, Naive Bayes, K-NN, and Support Vector Machine algorithms uses four approaches of Precision, Recall, F-Measure, and accuracy. The imbalanced dataset is divided into two parts with the composition of training data and testing data (80:20). Table 2 - 3 and Table 4 - 5 the training, data has precision, recall, f-measure, and accuracy.

Table 2 Dataset training with classifier Decision Tree and Naive Bayes

		Decision Tree		
Dataset	Precision	Recall	F-Measure	Accuracy
Abalone	97.70%	97.80%	97.50%	97.75%
Breast	82.30%	81.50%	79.80%	81.48%
E.coli	94.90%	93.40%	93.70%	93.36%
Robot	99.90%	99.90%	99.90%	99.91%
Yeast	94.30%	94.30%	94.10%	94.32%

Table 3 Dataset training with classifier Naive Bayes

		Naïve Bayes		
Dataset	Precision	Recall	F-Measure	Accuracy
Abalone	92.40%	82.70%	86.60%	82.73%
Breast	83.10%	64.20%	64.20%	64.20%
E.coli	87.60%	84.80%	85.60%	84.77%
Robot	80.00%	80.40%	80.20%	80.37%
Yeast	87.30%	87.50%	87.40%	87.46%

Table 4 Dataset training with classifier K-NN

		K-NN		
Dataset	Precision	Recall	F-Measure	Accuracy
Abalone	96.00%	95.90%	94.70%	95.85%
Breast	91.70%	91.40%	91.10%	91.36%
E.coli	94.80%	94.90%	94.80%	94.92%
Robot	95.00%	95.00%	95.00%	94.99%
Yeast	91.20%	91.50%	91.20%	91.53%

Table 5 Dataset training with classifier Support Vector Machine

	Support Vector Machine			
Dataset	Precision	Recall	F-Measure	Accuracy
Abalone	89.30%	94.50%	91.80%	94.47%
Breast	81.60%	79.00%	76.00%	79.01%
E.coli	89.00%	89.50%	89.00%	89.45%
Robot	76.60%	79.60%	74.20%	79.57%
Yeast	85.70%	86.40%	84.40%	86.36%

The Evaluation of four algorithms (Decision Tree, Naïve Bayes, K-NN, and Support Vector Machine) aims to determine the results of precision, recall, f-measure, and accuracy. Oversampling in the training dataset is done to make synthetic data into balanced data (figure 4 and 5).

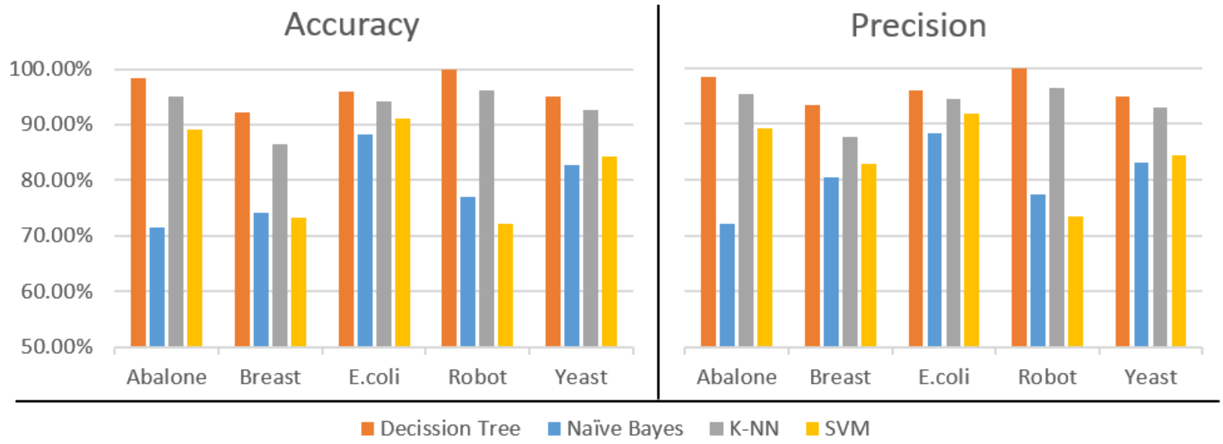

Figure 4. Result accuracy and precision

The imbalanced dataset obtained an accuracy evaluation (**figure 6**) with a value of 96.30% in the decision tree algorithm, K-NN 92.95%, Support Vector Machine 82%, and 78.74% in the Naïve Bayes classifier. 96.57% is the result of evaluating the precision of the decision tree classifier, Naïve Bayes 80.32%, for K-NN 93.38%, while for Support Vector Machine itself has a precision of 84.36% (figure 6).

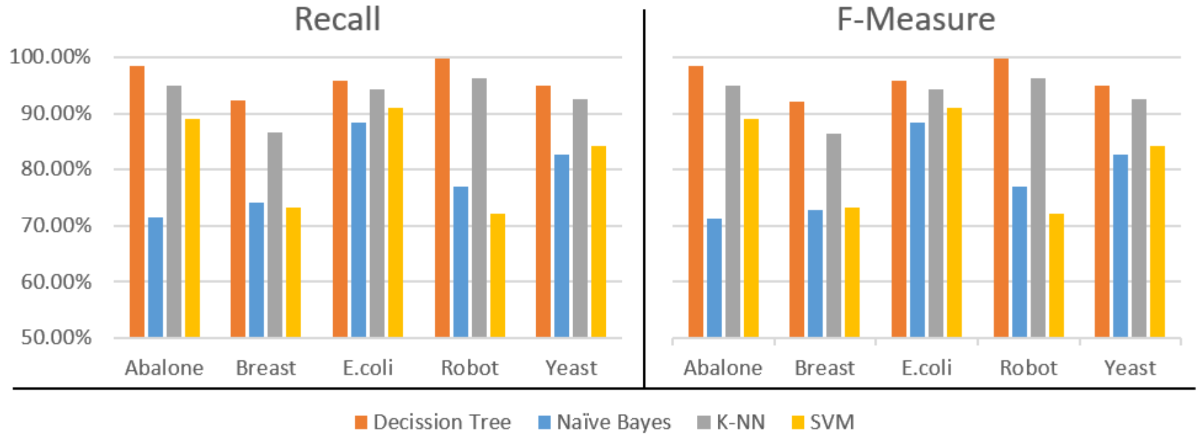

Figure 5. Result recall and F-Measure

Figure **7** is the result of the classifier evaluation conducted after the imbalanced dataset becomes balance by adding synthetic data to the minor class. Recall and F-Measure in the decision tree classifier are 96.31% and 93.30%. Naïve Bayes 78.74% and 78.38%, in K-NN, obtained 92.94% and 92.92%, and the Support Vector Machine has a recall and F-measure of 82% and 81.45%.

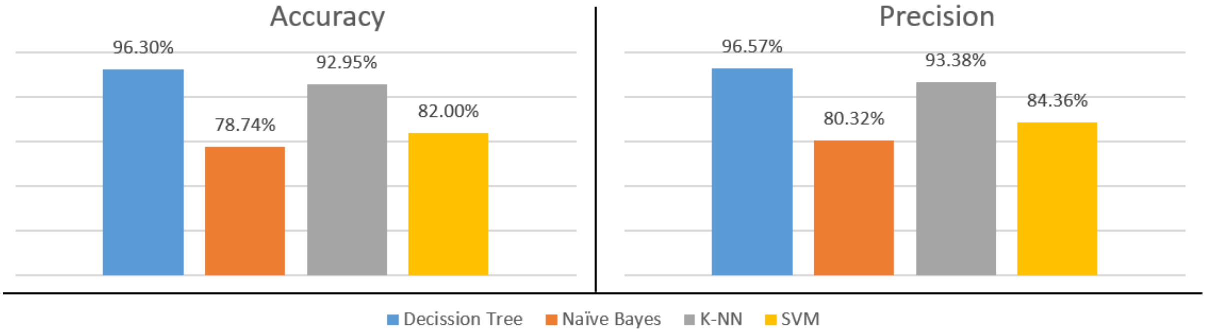

Figure 6. Average accuracy and precision

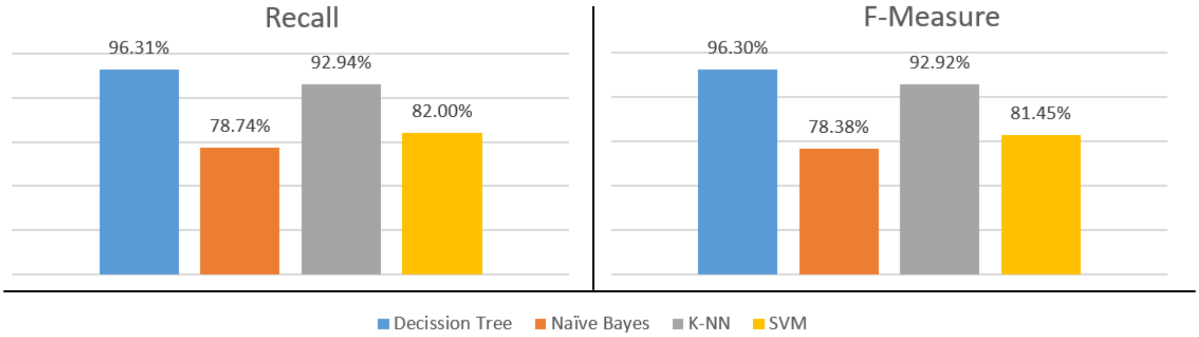

Figure 7. Average recall and F-Measure

5. Conclusion

In this research, the training data obtained an accuracy value of 93.73% K-NN, and the Naïve Bayes obtained an accuracy value of 79.90%. For Decision Tree test data has an accuracy value of 94.32, K-NN 92.67%, Support Vector Machine 85.61%, and Naïve Bayes 84.30%. After oversampling with MWMOTE on imbalanced data accuracy Decision Tree 96.30%, K-NN 92.95%, Support Vector Machine 82.00%, and Naïve Bayes 78.74%. That balanced data has less accuracy than balanced data with an average of 2-4%.

References

[1] M. J. Kim, D. K. Kang, and H. B. Kim, "Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1074–1082, 2015.
[2] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE - Majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 405–425, 2014.
[3] K. Napierała, "Improving Rule Classifiers For Imbalanced Data," no. October, 2012.
[4] P. Phoungphol, "A Classification Framework for Imbalanced Data," pp. 1–94, 2013.
[5] S. Jayasree and A. A. Gavya, "Classification of Imbalance Problem by MWMOTE and SSO," *Ijmtes*, pp. 1–4, 2015.
[6] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the oversampling of different classes

and types of examples in multi-class imbalanced datasets," *Pattern Recognit.*, vol. 57, pp. 164–178, 2016.

[7] A. Mellor, S. Boukir, A. Haywood, and S. Jones, "Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin," *ISPRS J. Photogramm. Remote Sens.*, vol. 105, pp. 155–168, 2015.

[8] J. Gong and H. Kim, "RHSBoost: Improving classification performance in imbalance data," *Comput. Stat. Data Anal.*, vol. 111, pp. 1–13, 2017.

[9] I. Fakhruzi, "An artificial neural network with bagging to address imbalance datasets on clinical prediction," *2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018*, vol. 2018-Janua, no. 1, pp. 895–898, 2018.

[10] A. B. Mohammed, "Decision Tree , Naïve Bayes and Support Vector Machine Applying on Social Media Usage in NYC / Comparative Analysis Decision Tree , Naïve Bayes and Support Vector Machine Applying on Social Media Usage in NYC / Comparative Analysis," *Tikrit J. Pure Sci.*, vol. 22, no. January 2017, pp. 94–99, 2019.

[11] F. I. Komputer and U. D. Nuswantoro, "Penyakit Stroke Dengan Klasifikasi Data Mining Pada," p. 7, 2011.

[12] Y. E. Kurniawati, A. E. Permanasari, and S. Fauziati, "Adaptive Synthetic-Nominal (ADASYN-N) and Adaptive Synthetic-KNN (ADASYN-KNN) for Multiclass Imbalance Learning on Laboratory Test Data," *Proc. - 2018 4th Int. Conf. Sci. Technol. ICST 2018*, vol. 1, pp. 1–6, 2018.

[13] A. Verasius and D. Sano, "Comparison of Prediction Accuracy Between Decision Tree , Naïve Bayes and K-Nn on Web Phising," pp. 380–384, 2018.

[14] M. Nishom, "Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square," *J. Inform. J. Pengemb. IT*, vol. 4, no. 1, pp. 20–24, 2019.

[15] G. De Leonardis *et al.*, "Human Activity Recognition by Wearable Sensors : Comparison of different classifiers for real-time applications," *MeMeA 2018 - 2018 IEEE Int. Symp. Med. Meas. Appl. Proc.*, vol. 3528725544, pp. 1–6, 2018.

[16] E. Wijaya, "Implementation Analysis of GLCM and Naive Bayes Methods in Conducting Extractions on Dental Image," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 407, no. 1, 2018.

[17] S. Guo, D. Guo, L. Chen, and Q. Jiang, "A centroid-based gene selection method for microarray data classification," *J. Theor. Biol.*, vol. 400, pp. 32–41, 2016.

[18] J. A. S. Almeida, L. M. S. Barbosa, A. A. C. C. Pais, and S. J. Formosinho, "Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering," *Chemom. Intell. Lab. Syst.*, vol. 87, no. 2, pp. 208–217, 2007.

[19] C. Seiffert, T. M. Khoshgoftaar, and J. Van Hulse, "Hybrid sampling for imbalanced data," *Integr. Comput. Aided. Eng.*, vol. 16, no. 3, pp. 193–210, 2009.

[20] A. M. Mahmood, "Class Imbalance Learning in Data Mining – A Survey," *Int. J. Commun. Technol. Soc. Netw. Serv.*, vol. 3, no. 2, pp. 17–36, 2015.

[21] J. Alcalá-Fdez *et al.*, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Mult. Log. Soft Comput.*, vol. 17, no. 2–3, pp. 255–287, 2011.

[22] A. Ashari, I. Paryudi, and A. Min, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 11, pp. 33–39, 2013.

[23] T. Pranckevičius and V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," *Balt. J. Mod. Comput.*, vol. 5, no. 2, pp. 221–232, 2017.