Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE

Anthony Anggrawan *, Hairani Hairani, and Christofer Satria

***Abstract*—Student graduation accuracy is one of the indicators of the success of higher education institutions in carrying out the teaching and learning process and as a component of higher education accreditation. So it is not surprising that building a system that can predict or classify students graduating on time or not on time is necessary for universities to monitor the exact number of students graduating on time using educational technology. Unfortunately, educational technology or machine learning with data mining approaches is less accurate in classifying classes with unbalanced data. Therefore, this research purpose is to build a machine learning system that can improve classification performance on unbalanced class data between students who graduate on time and graduate late. This study applies the Synthetic Minority Oversampling Technique (SMOTE) method to improve the classifying performance of the Support Vector Machine (SVM) data mining method. The results of the study concluded that using the SMOTE method increased the accuracy, precision, and sensitivity of the SVM method in classifying class data of unbalanced student graduation times. The SVM performance score rises by 3% for classification accuracy, 8% for classification precision, and 25% for classification sensitivity.**

***Index Terms*—Classification, educational technology, machine learning, data mining, Support Vector Machine (SVM), Synthetic Minority Oversampling Technique (SMOTE).**

I. Introduction

Although educational information technology supports learning today [1–4], graduation and timeliness of graduation are different achievements for all students [5]. Many factors affect the timely completion of studies for students [6–8]. Statistics show that the average speed/punctuality of student graduation is not the same time [9], specifically, there is an imbalance between students who are on time or graduating quickly and those who are not on time or late for graduation [8]. Meanwhile, the graduation rate on time is one indicator of the success of learning in higher education [6, 10] and is one of the elements of the assessment of higher education accreditation in Indonesia [10], in addition to other elements that indicate the success of higher education [11–13]. Therefore, building a system that can predict or classify the accuracy of student graduation is one way for universities to monitor the certainty of student graduation precisely and not on time [10]. However, there are obstacles encountered in building an application system in classifying the accuracy of graduation, namely the accuracy of the system constructed especially on unbalanced class data between the number of students who graduate on time and do not graduate on time.

According to information from the Ministry of Higher Classifying, unbalanced class data is a significant problem in machine learning and data mining. Because, after all, causes inaccuracy in classification is the imbalance of class data [14, 15]. It happened because the imbalance distribution of class data causes biased classifier performance due to misclassifying the minority class or minority classes not being considered in the overall classification results [16]. Worse, machine learning methods ignore unbalanced data, so machine learning training with unbalanced class data negatively impacts machine learning performance [17]. As a result, machine learning models perform poorly in the minority class [18]. In other words, the classification method does not achieve maximum performance when applied to unbalanced class data [18, 19]. That is why the problem of unbalanced data sets gets special attention in machine learning and research related to machine learning [14, 16] and has become a hot issue in data mining [20, 21]. In short, classification research on unbalanced classes is essential; moreover, a class imbalance is inherent in much of the natural world [22] and not just in machine learning [17].

In essence, the classification model is a popular data mining or machine learning model [23–25] and has its application in various fields of science [26]. The classification model is a predictive learning model through training data on the data set to identify the pattern of relationships between attributes and classes in the data set [27, 28]. Predicting is not an easy task [13, 29]; difficulties arise due to considering several criteria as the basis for prediction or decision-making [13, 30]. Therefore, previous researchers emphasized that what often happens is inaccuracy in making decisions [29]. That is why there is a need for a system that can assist in predicting with reasonable accuracy the results. Machine learning can predict accurately [25]. Machine learning has artificial intelligence in carrying out its jobs. Artificial intelligence [25, 31] is today's learning technology widely used for various roles [31]. Through machine learning, it is possible to uncover hidden patterns in big data and classify them [32].

Although, there are several classification methods: SVM, Random Forest, Naive Bayes, Decision Tree, and others [2], [27, 33]. However, SVM is a widely known method used for classification [34]. Each classification method has a different classification accuracy level. At the same time, inaccurate classifying of events results in errors in identifying particular patterns from the data set. SVM is a classification method used as a training system for linear learning machines [35]. As a result, machine learning can accurately perform classification [25]. However, according to López and

Manuscript received July 23, 2022; revised September 3, 2022; accepted September 5, 2022.

Anthony Anggrawan is with Information Technology Education Department, Bumigora University, Indonesia.

Hairani Hairani is with Computer Science Department, Bumigora University, Indonesia (e-mail: hairani@universitasbumigora.ac.id).

Christofer Satria is with Visual Design Communication, Bumigora University, Indonesia (e-mail: chris@universitasbumigora.ac.id).

*Correspondence: anthony.anggrawan@universitasbumigora.ac.id

doi: 10.18178/ijiet.2023.13.2.1806

Fernández *et al*. [36], SVM machine learning and decision trees are unsuitable for producing good performance on unbalanced class data; therefore, it is not surprising that the imbalance of data on class attributes encourages many researchers to study it [19, 37, 38]. For this reason, this study aims to improve the performance of predictions or classification of the timeliness of graduating students by using SMOTE and SVM methods. Furthermore, to prove an increase in the accuracy of classifying or predicting classes on time for graduation, this study compared the performance results between the SVM method combined with the SMOTE method and the SVM method without the combination with the SMOTE method.

SMOTE is a resampling method [39] that can improve classification performance on unbalanced data, especially when combined with other methods [40]. However, the question is whether the application of SMOTE can improve the predictive performance of SVM data mining methods on unbalanced class data on the student graduation timeliness dataset? Also, how much precision/accuracy/sensitivity is the application of SMOTE in improving the classification or predictive performance of the SVM data mining method on unbalanced data from the class on the timeliness of graduation students? This research proves it.

Further discussion in this manuscript is as follows. The second section deals with related work. The third section describes the research methodology. The fourth section explains the results and discussion of the research. Finally, the fifth section is the conclusions that discusses conclusions, updates, and suggestions for further investigation.

II. Related Work

Some of the latest related works of previous research are as follows.

Bartosz Krawczyk discusses the challenges open to researchers and future research directions for unbalanced data class [14]. The previous research differs from the research in this article not only in the research method but also in the research objectives. The previous research was a literature study review paper on unbalanced data classes. In contrast, the research in this article is an experimental study to improve the prediction performance of unbalanced class data from data on student graduation timeliness.

Elreedy and Atiya [41] presented an analysis of the SMOTE method. This last study introduced how to overcome the classification problem of unbalanced data in the minority class by generating additional data from the minority class using SMOTE. So this previous research has a different objective (focus) compared to the research in this article. The previous research describes how SMOTE makes unbalanced class data into balance class data. In contrast, this article's research improves the SVM method's performance in classifying unbalanced data from student graduation accuracy data. In the meantime, Johnson and Khoshgoftaar [22] surveyed the literature on using deep learning methods to address class data unbalances. The previous research was survey research to overcome unbalanced class data with deep learning methods. In contrast to the research in this article is a trial study of the application of the SMOTE method to improve the accuracy of the SVM method classification in survey research to overcome unbalanced class data with deep dealing with unbalanced class data.

Patel and Singh Rajput *et al*. [16] reviewed the classification of unbalanced data on wireless sensor networks. However, this previous research has different objectives, objects, and methods compared to the study conducted in this article. Kumar and Bhatnagar *et al.* [17] presented various approaches to classifying unbalanced data sets. The main difference lies in the research methods and objectives between the previous research and the research in this article. The previous research was a review study related to the unbalanced class data classification approach. In contrast, the research in this article is an experimental study to improve the classification performance of the SVM data mining method.

Meanwhile, Wang and Shen *et al*. [20] proposed the use of the SMOTE method to improve the classification results of the Random Forest classification method for several data sets. However, this previous study focused on enhancing classification performance using SMOTE on the Random Forest data mining method and not on student pass accuracy data. In contrast, this research focused on improving classification performance with SMOTE on the SVM data mining method on unbalanced student pass accuracy data.

Huang and Dai [26] reviewed the class data imbalance in the Decision Tree method. The difference with this article is in the research objectives and research methods. Previous research focused on discussing unbalanced class data on the Decision Tree method. In contrast, the research in this article focuses on testing classification performance to unbalanced class data on the student graduation timeliness data set on the SVM method.

In contrast, Wang and Xu *et al*. [42] proposed a scheme that can decide the composition of the training data for federated learning to reduce the impact of class data imbalance. This previous study proposed a method for detecting class data imbalances in federated learning and reducing the effect of class data imbalance, in contrast to the research in this article, which focuses on applying the SMOTE method to improve prediction accuracy on unbalanced class data in the SVM method.

Zheng and Jin *et al*. [43] investigated the performance effect of unbalanced class data and training data measures for classifiers. This previous research is an empirical study on the Naive Bayes, logistic regression, and Tree methods. Previous research compared balanced and unbalanced data to measure the accuracy of data mining methods; in contrast to this article's research, the mining method improves performance (accuracy, precision, and sensitivity) by applying the SMOTE method to the mining method. The research in this article then compares the performance of the data mining method between those implementing the SMOTE method and those not using the SMOTE method.

The review of several prior research-related works confirms that the study of this article differs from previous associated works. The findings of this study help reveal the impact of increasing classification accuracy arising from the application of the SMOTE method to the data set on the imbalance in the timeliness of students' graduation in the SVM method. The novelty of this study lies in improving the classification performance or prediction of unbalanced class data on student graduation timeliness which previous researchers have never done.

TABLE I: COMPARISON OF THIS ARTICLE'S WORK WITH SOME PREVIOUS RELATED WORKS

Research By	Type of Research	Method Used		Performance Testing			Research Object	Research Data / Data Set
		SVM	SMOTE	Accuracy	Precision	Sensitivity		
Bartosz Krawczyk [14]	Review	No	No	Yes	Yes	Yes	Reviewing methods for dealing with unbalanced class data problems on the Decision Tree method	Various data sets depending on the reviewed article, for example, Behavior, Cancer malignancy grading, Hyperspectral data, and others
Dina Elreedy *et al*.[41]	Theoretical and experimental	No	Yes	Yes	No	No	Test the classification accuracy using SMOTE on K-nearest neighbors (KNN) method	Multivariate Gaussian distribution data
Justin M. Johnson *et al*. [22]	Survey	No	No	No	No	No	Surveying existing deep learning techniques to overcome unbalanced class data	Various data sets depending on the surveyed article, for example, CIFAR-10, Public cameras, Building changes, and others
Harshita Patel *et al*. [16]	Review	No	No	No	No	No	Troubleshooting data imbalance issues of a wireless sensor network on the KNN method	No specifically mention
Pradeep Kumar *et al*. [17]	Review	Yes	No	No	No	No	Reviewing various data imbalance issues and learning strategies and algorithms from the Random Forest, KNN, Decision Tree, Neural Network, Naive Bayes, and SVM classification techniques.	No specifically mention (except imbalanced data)
Shujuan Wang *et al*. [20]	Experimental	No	Yes	Yes	No	No	Improving classification results Random Forest method for multiple data sets	Pima, WDBC, WPBC, Ionosphere, and Breast-cancer-Wisconsin
Cui Yin Huang *et al*.[26]	Experimental	No	Yes	Yes	Yes	Yes	Reviewing the class data imbalance in the Decision Tree method	Yeast, Glass, Cleveland, and Vehicle
Lixu Wang *et al*. [40]	Experimental	No	No	No	No	No	Propose a scheme to decide the composition of training data to reduce the impact of class data imbalance	Clients or server data
Wanwan Zheng *et al*. [41]	Experimental	No	No	No	No	No	Investigating the performance effects of unbalanced class data and training data measures for classifiers in the Naive Bayes, logistic regression, and Tree methods	Ozone. Kc1, Scene, Gesture, Cpu_act, Waveform-5000, Spambase, and Madelone
Our/this research	Experimental	Yes	Yes	Yes	Yes	Yes	Test the performance of the SVM method classification on the timeliness of graduating students	Student Graduation Data

In other words, the advantage of this research is that it is an experimental study on the imbalance of data on student graduation timeliness with SMOTE in SVM that other researchers have not studied. Table I shows the comparison between the previous related studies and this article.

III. RESEARCH METHODOLOGY

This study uses data mining stages, as shown in Fig. 1.

A. Data Collection

Data collection was carried out at Bumigora University. The data set was taken from graduation data for undergraduate students for the 2019-2021 academic years, totaling 265 data and having eight attributes. The attributes of this research data set are shown in Table II. The data used as machine learning training data in this study is the achievement index (IP) data from student graduation data for six semesters who have completed their studies. Machine learning is helpful for systematically predicting which students will graduate on time and who will be late for graduation based on variations in the 6-semester achievement index value, which has a decimal value variation of 0.0 to 4.0. Students with a good to excellent achievement index have a minimum achievement index of 2.0. Research data shows that not always students who excel and are very good will definitely graduate on time (see the data set in Table III). Machine learning that implements data mining methods has intelligence that can reveal hidden patterns in big data [32] and can predict with high accuracy [25]. In other words, machine learning has the intelligence to predict students who have completed their studies up to semester six whether these students will graduate on time or not. The sample data for students' graduation is shown in Table III.

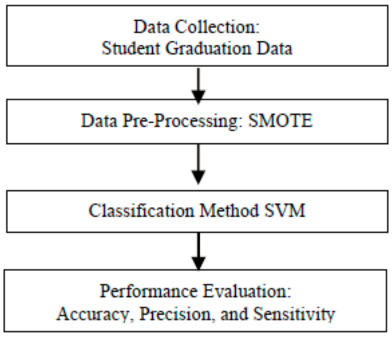

Fig. 1. Research stages.

TABLE II: STUDENT GRADUATION DATASET ATTRIBUTES

No	Attribute Name	Information	Data Type
1	JK	Gender	Nominal (Male, Female)
2	IPS 1	Semester 1 IP	Numerical
3.	IPS 2	Semester 2 IP	Numerical
4.	IPS 3	Semester 3 IP	Numerical
5.	IPS 4	Semester 4 IP	Numerical
6.	IPS 5	Semester 5 IP	Numerical
7.	IPS 6	Semester 6 IP	Numerical
8.	Graduation Status	Class	Nominal (On Time, Not On Time)

TABLE III: STUDENT GRADUATION DATASET

No	JK	IPS1	IPS2	…	IPS6	Status Graduation
1	F	3.06	3.16	…	3.17	On-Time
2	F	3.41	3.43	…	3.44	On-Time
3	M	2.43	2.61	…	2.67	Not On Time
4	F	3.5	3.53	...	3.53	On-Time
5	M	2.07	2.22	…	2.32	Not On Time
6	F	3.42	2.85	…	3.5	On-Time
7	M	3.33	3.28	…	3.15	Not On Time
8	F	2.83	2.05	…	2.66	Not On Time
9	M	2.94	2.21	…	3.1	Not On Time
10	M	2.56	2.0	…	2.68	Not On Time
..	...	……	……	…	…..	……………
264	M	2.69	1.85	…	2.5	Not On Time
265	F	2.22	1.83	…	2.21	Not On Time

B. Data Pre-processing

Data Pre-processing is one of the crucial stages in data mining to improve the quality of data sets. This study deals with unbalanced data contained in student graduation data sets. The dataset used has 171 data classes that are not on time and 94 data on time. The algorithm used to handle unbalanced data in the dataset is SMOTE (Synthetic Minority Oversampling Technique).

Attributes with categorical data types are converted to numeric data types before the oversampling process using SMOTE. The gender attribute has a categorical data type with categories 'L' and 'P', so the category 'L' becomes 0, and 'P' becomes 1.

SMOTE is one of the most commonly used oversampling methods to solve the problem of data distribution imbalance in machine learning modeling. SMOTE aims to balance the distribution of classes by increasing the number of minority classes by synthesizing data for oversampling purposes [29]. Creating new data for the minority class uses Eq. (1).

$$Y' = Y^i + (Y^j - Y^i) * \gamma \tag{1}$$

Y': is to hold the result of the new data. Y^i : represents the minority class. Y^j : is a randomly selected value from the k-nearest neighbors of the minority class Y^i, and γ: is a randomly selected value in a random vector with a range of 0 to 1 [44]. SMOTE generates new synthesis training data by linear interpolation for the minority class. Synthesis training data is generated by randomly selecting one or more of the k-nearest neighbors for each sample in the minority class, as shown in Fig. 2.

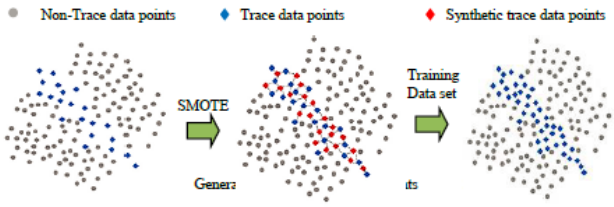

Fig. 2. Synthetic minority oversampling technique (SMOTE) algorithm working process [45].

C. Classification Method

The realization of classification data mining using data mining methods or machine learning algorithms involves two data sets: the first is the dataset for training, and the second is for testing. Each item set involves the attributes and categories of each training attribute with a specific target value.

This study uses the SVM data mining method to classify student graduation. Before classification, the dataset is first divided into training and testing data using 10-fold cross-validation, divided into 10 data groups using python tools.

The SVM data mining method is a supervised learning classification method aiming to find the optimal hyperplane by maximizing the distance or margin between data classes using Eq. (2) [46].

$$h(x) = w^T x + b \tag{2}$$

$$w^T x_i + b \geq +1 \text{ when } y_i = +1 \tag{3}$$

$$w^T x_i + b \leq -1 \text{ when } y_i = -1 \tag{4}$$

w is a weight vector; *x* is the input vector; *b* is biased.

The SVM method works not only on linear but also on nonlinear data. The technique uses two approaches to transform nonlinear data into linear data: soft margin hyperplane and feature space. The soft margin hyperplane approach in converting nonlinear data into linear ones is with the slack ξ variable formulation, as shown in equations (5) and (6). The parameters used in the SVM method are kernel RBF, C = 5, gamma = 2, and toll = 0.0001. The use of these parameters is the best combination of parameters for the SVM method on the dataset used based on the results of hyperparameter tuning using the Grid search technique to improve accuracy.

$$x_i\, w_i + b \geq 1 - \xi \text{ for } y_i = \text{class } 1 \tag{5}$$

$$x_i\, w_i + b \leq -1 + \xi \text{ for } y_i = \text{class } 24 \tag{6}$$

D. Performance Evaluation

Evaluation (testing) of performance uses a confusion matrix. The Confusion Matrix helps calculate the amount of data classified as true and false, as shown in Table IV.

TABLE IV: CONFUSION MATRIX

Actual	Prediction	
	On-time	Not on time
On-time	TP	FN
Not On time	FP	TN

The formula for calculating accuracy, precision, and sensitivity is as follows: [28, 47]

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (7)$$

$$Precision = \frac{TP}{TP + FP} \quad (8)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (9)$$

True Positive (TP) is a class on time that is correctly predicted. False Positive (FP) is a class that is not on time but is predicted to be on time. True Negative (TN) is an incorrectly predicted class on time. False Negative (FN) is a class that is on time but is predicted not to be on time.

Accuracy states the closeness of the measurement results to the actual value, while precision shows how close the difference in the measurement results is on repeated measurements. On the other hand, sensitivity states the level of success in retrieving information. The accuracy measurement is based on the ratio between the correct predictions (positive and negative) with the overall data. In contrast, precision measurements are based on the percentage of true positive predictions compared to overall positive predicted outcomes. Meanwhile, the recall measurement is based on the ratio of true positive predictions compared to the general actual positive data.

IV. RESULT AND DISCUSSION

This research starts from the stages of data collection, data pre-processing, classification, and performance testing. The data used in this study is the graduation data of students. Pre-processing this research uses the SMOTE algorithm to deal with class imbalances in the graduation data used. The results of comparing the original data with the data from SMOTE are shown in Fig. 3.

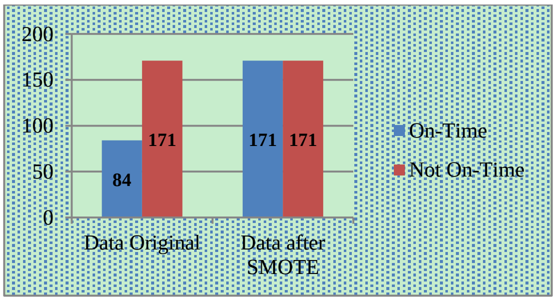

Fig. 3. The results of the comparison of the original data with the data from the SMOTE.

TABLE V: CONFUSION MATRIX RESULT OF SVM METHOD

Actual	Predicted	
	On-Time	Not On Time
On-Time	50	44
Not On Time	23	148

Table V shows the results of testing the SVM method with a confusion matrix using 10-fold cross-validation. Meanwhile, Table VI shows the results of the Confusion Matrix with the SVM method and the SMOTE method.

TABLE VI: CONFUSION MATRIX RESULT OF SVM AND SMOTE METHODS

Actual	Predicted	
	On-Time	Not On Time
On-Time	134	37
Not On Time	43	128

Table VII and Fig. 2 show an increase in the performance of the SVM method with SMOTE based on accuracy, precision, and sensitivity. Without SMOTE, the SVM method has 74% accuracy, 68% precision, and 53% sensitivity. While using SMOTE, the SVM method has an accuracy of 77%, 76% precision, and a sensitivity of 78%. In other words, the SVM performance score using SMOTE for accuracy increased by 3%, precision increased by 8%, and sensitivity increased by 25%. Thus, this study concludes that using the SMOTE method improves the accuracy, precision, and sensitivity of the SVM method in managing unbalanced class category data. Furthermore, using SMOTE sampling reduces the skewness of the data distribution to improve the performance of the classification method used [48, 49].

TABLE VII: PERFORMANCE RESULT OF CLASSIFICATION METHOD

Method	Accuracy	Precision	Sensitivity
SVM	74%	68%	53%
SVM with SMOTE	77%	76%	78%

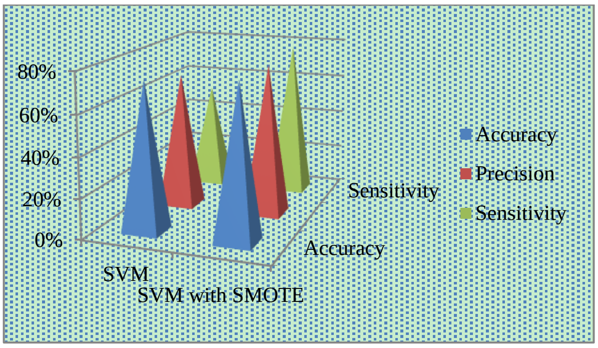

Fig. 4. Performance result of classification method.

V. CONCLUSION

The results of this study prove that the SMOTE method helps improve the performance of the accuracy, precision, and sensitivity of the SVM data mining method or the SVM machine learning algorithm in managing unbalanced student graduation time data. Furthermore, the results show the novelty of the discovery, namely the SVM performance score using SMOTE to reach 3% for the accuracy of the classification results of unbalanced class data on student graduation timeliness and up to 25% for the sensitivity of the classification results of unbalanced class data on student graduation timeliness. Meanwhile, using SMOTE, the SVM performance score increased its precision by 8% in predicting students' on-time and not on-time graduation.

Further research needs to conduct SMOTE testing for other machine learning algorithms and research with more complex data sets to meet SMOTE needs. In addition, it is necessary to further develop the results of this research by building a Web or Cloud-based application program and testing its implementation on users. Finally, further research can also combine several ensemble learning-based methods with SMOTE to get better accuracy with other datasets.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

All authors participated in completing the research and writing of this article. The level of roles and tasks of research work is the basis that places each author as a correspondent author, first author, and second author.

REFERENCES

[1] S. Smith, D. Cobham, and K. Jacques, "The use of data mining and automated social networking tools in virtual learning environments to improve student engagement in higher education," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 4, pp. 263–271, 2022.

[2] C. Teoh, S. Ho, K. S. Dollmat, and C. Tan, "Ensemble-Learning techniques for predicting student performance on video-based learning," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 8, pp. 741–745, 2022.

[3] A. Anggrawan, C. Satria, Mayadi, and N. G. A. Dasriani, "Reciprocity effect between cognitive style and mixed learning method on computer programming skill," *J. Comput. Sci.*, vol. 17, no. 9, pp. 814–824, 2021.

[4] A. Anggrawan, "Interaction between learning preferences and methods in face-to-face and online learning," *ICIC Express Lett.*, vol. 15, no. 4, pp. 319–326, 2021.

[5] V. Bocsi *et al.*, "The discovery of the possible reasons for delayed graduation and dropout in the light of a qualitative research study," *J. Adult Learn. Knowl. Innov.*, vol. 3, no. 1, pp. 27–38, 2019.

[6] N. Mohammad Suhaimi, S. Abdul-Rahman, S. Mutalib, N. H. Abdul Hamid, and A. Md Ab Malik, "Predictive model of graduate-on-time using machine learning algorithms," *Commun. Comput. Inf. Sci.*, vol. 1100, no. September, pp. 130–141, 2019.

[7] A. Anggrawan, A. H. Yassi, C. Satria, B. Arafah, and H. M. Makka, "Comparison of online learning versus face to face learning in English grammar learning," in *Proc. 5th International Conference on Computing Engineering and Design*, 2018, pp. 1–4.

[8] T. Ojha, G. L. Heileman, M. Martinez-Ramon, and A. Slim, "Prediction of graduation delay based on student performance," in *Proc. International Joint Conference on Neural Networks*, 2017-May, pp. 3454–3460.

[9] J. M. Aiken, R. de Bin, M. Hjorth-Jensen, and M. D. Caballero, "Predicting time to graduation at a large enrollment American university," *PLoS One*, vol. 15, no. 11, pp. 1–28, Nov. 2020.

[10] D. Kurniawan, A. Anggrawan, and H. Hairani, "Graduation prediction system on students using C4.5 algorithm," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 19, no. 2, pp. 358–365, 2020.

[11] A. Anggrawan, "Percentage of effect of blended learning madel on learning outcome," in *Proc. 4th International Conference on Informatics and Computing*, 2019.

[12] M. Cheon, O. Lee, C. Mun, and H. Ha, "Factors affecting academic achievement in SW education," *Int. J. Inf. Educ. Technol.*, vol. 12, no. 4, pp. 333–338, 2022.

[13] A. Anggrawan, Mayadi, C. Satria, and L. G. R. Putra, "Scholarship recipients recommendation system using AHP and moora methods," *Int. J. Intell. Eng. Syst.*, vol. 15, no. 2, pp. 260–275, 2022.

[14] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[15] D. S. Rajput, R. S. Thakur And, and G. S. Thakur, "A computational model for knowledge extraction in uncertain textual data using karnaugh map technique," *Int. J. Comput. Sci. Math.*, vol. 7, no. 2, pp. 166–176, 2016.

[16] H. Patel, D. Singh Rajput, G. Thippa Reddy, C. Iwendi, A. Kashif Bashir, and O. Jo, "A review on classification of imbalanced data for wireless sensor networks," *Int. J. Distrib. Sens. Networks*, vol. 16, no. 4, pp. 1–15, 2020.

[17] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of imbalanced data: Review of methods and applications," in *Proc. IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1099, no. 1, p. 012077.

[18] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, 2017.

[19] D. Devi, S. K. Biswas, and B. Purkayastha, "A review on solution to class imbalance problem: Undersampling approaches," in *Proc. International Conference on Computational Performance Evaluation*, 2020, pp. 626–631.

[20] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021.

[21] D. S. Rajput, "Review on recent developments in frequent itemset based document clustering, its research trends and applications," *Int. J. Data Anal. Tech. Strateg.*, vol. 11, no. 2, pp. 176–195, 2019.

[22] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, 2019.

[23] D. Prasad, S. Kumar Goyal, A. Sharma, A. Bindal, and V. Singh Kushwah, "System model for prediction analytics using k-nearest neighbors algorithm," *J. Comput. Theor. Nanosci.*, vol. 16, no. 10, pp. 4425–4430, 2019.

[24] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, "Early dropout prediction using data mining: A case study with high school students," *Expert Syst.*, vol. 33, no. 1, pp. 107–124, 2016.

[25] A. Anggrawan, C. Satria, C. K. Nuraini, Lusiana, N. G. A. Dasriani, and Mayadi, "Machine learning for diagnosing drug users and types of drugs used," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 111–118, 2021.

[26] C. Y. Huang and H. L. Dai, "Learning from class-imbalanced data: Review of data driven methods and algorithm driven methods," *Data Sci. Financ. Econ.*, vol. 1, no. 1, pp. 21–36, 2021.

[27] A. Yosipof, R. C. Guedes, and A. T. García-Sosa, "Data mining and machine learning models for predicting drug likeness and their disease or organ category," *Front. Chem.*, vol. 6, no. May, pp. 1–11, 2018.

[28] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The abstract of thesis classifier by using naive bayes method," in *Proc. International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management*, 2021, no. August, pp. 312–315.

[29] D. Laureiro-Martínez and S. Brusoni, "Cognitive flexibility and adaptive decision-making: Evidence from a laboratory study of expert decision-makers," *Strateg. Manag. J.*, vol. 39, no. 4, pp. 1031–1058, 2018.

[30] P. H. Santos, S. M. Neves, D. O. Sant'Anna, C. H. Oliveira, and H. D. Carvalho, "The analytic hierarchy process supporting decision making for sustainable development: An overview of applications," *J. Clean. Prod.*, vol. 212, pp. 119–138, 2019.

[31] K. Shankar, A. R. W. Sait, D. Gupta, S. K. Lakshmanaprabu, A. Khanna, and H. M. Pandey, "Automated detection and classification of fundus diabetic retinopathy images using synergic deep learning model," *Pattern Recognit. Lett.*, vol. 133, pp. 210–216, 2020.

[32] G. Schneider, "Automating drug discovery," *Nat. Rev. Drug Discov.*, vol. 17, no. February, pp. 97–113, 2018.

[33] P. Kaur, M. Singh, and G. S. Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector," *Procedia Comput. Sci.*, vol. 57, pp. 500–508, 2015.

[34] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 13, pp. 6308–6325, 2020.

[35] J. C. Moreano and N. B. L. S. Palomino, "Global facial recognition using gabor wavelet, support vector machines and 3d face models," *J. Adv. Inf. Technol.*, vol. 11, no. 3, pp. 143–148, 2020.

[36] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci. (Ny).*, vol. 250, pp. 113–141, 2013.

[37] A. Onan, "Consensus clustering-based undersampling approach to imbalanced learning," *Sci. Program.*, vol. 2019, pp. 1–14, 2019.

[38] X. Wang, H. Wang, D. Wu, Y. Wang, and R. Zhou, "A fuzzy consensus clustering based undersampling approach for class imbalanced learning," *ACM Int. Conf. Proceeding Ser.*, vol. December, pp. 133–137, 2019.

[39] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. Sept. 28, pp. 321–357, 2002.

[40] V. López, A. Fernández, M. J. Del Jesus, and F. Herrera, "A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline datasets," *Knowledge-Based Syst.*, vol. 38, pp. 85–104, 2013.

[41] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019.

[42] L. Wang, S. Xu, X. Wang, and Q. Zhu, "Addressing class imbalance in federated learning," in *Proc. 35th AAAI Conference on Artificial Intelligence*, 2021, vol. 11B, pp. 10165–10173.

[43] W. Zheng and M. Jin, "The effects of class imbalance and training data size on classifier learning: an empirical study," *SN Comput. Sci.*, vol. 1, no. 2, pp. 1–11, 2020.

[44] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems Tuanfei," *Pattern Recognit.*, vol. 72, no. December, pp. 327–340, 2017.

[45] J. Chen, H. Huang, A. G. Cohn, D. Zhang, and M. Zhou, "Machine learning-based classification of rock discontinuity trace : SMOTE oversampling integrated with GBT ensemble learning," *Int. J. Min. Sci. Technol.*, vol. 32, no. 2, pp. 309–322, 2022.

[46] B. Singh and B. C. S. Rai, "Analysis of support vector machine-based intrusion detection techniques," *Arab. J. Sci. Eng.*, vol. 45, no. July, pp. 2371–2383, 2019.

[47] A. Luque, A. Carrasco, A. Mart ń, and A. De, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognit.*, vol. 91, pp. 216–231, 2019.

[48] S. Mishra, P. K. Mallick, L. Jena, and G. S. Chae, "Optimization of skewed data using sampling-based preprocessing approach," *Front. Public Heal.*, vol. 8, no. July, pp. 1–7, 2020.

[49] N. Noorhalim, A. Ali, and S. M. Shamsuddin, "Handling imbalanced ratio for class imbalance problem using SMOTE," in *Proc. 3rd International Conference on Computing, Mathematics and Statistics*, 2019, pp. 19–30.

Anthony Anggrawan received his masters in master of engineering (M.T) from the 10 November Institute of Technology, Surabaya, Indonesia. After that, he earned his first doctoral degree (Ph.D.) in accounting information systems from Universiti Utara Malaysia. Then, he received his second doctoral degree (Dr.) from Hasanuddin University, Makassar, Indonesia, in the linguistics field. Finally, he earned his third doctorate in educational technology from the State University of Jakarta. He currently works as an associate professor in the Department of Information Technology Education as a lecturer, University Rector, and State Civil Apparatus. His research interests include educational technology, machine learning, online learning, data mining, and the internet of things. During this time, he is active as an article reviewer in several reputable international scientific journals.

Hairani Hairani obtained a bachelor's degree (S.Kom) in computer science from Islamic University of Indonesia, Yogyakarta, Indonesia, and a master's degree (M.Eng) in master of engineering from the Gajah Mada University, Yogyakarta, Indonesia. He currently serves as a lecturer in the Computer Science study program, Bumigora University, Mataram, Indonesia, and a member of the Institute for Research and Community Service. His research interests include data mining, machine learning and artificial intelligence.

Christofer Satria received a bachelor's degree (S.Sn) in visual communication design from Petra Christian University, Surabaya, Indonesia, and a master's degree (M.Sn) in visual communication design from the Indonesian Art Institute (ISI) Denpasar, Bali, Indonesia. He is currently a lecturer in the visual communication design study program at Bumigora University, Indonesia, and the head of the laboratory in photography, animation, and video. His research interests include animation learning media, video learning media, education method, data mining, and experimental design. He is currently pursuing a doctorate in the same area as his expertise.