Recursive Journal of Informatics
Vol. 1, No. 1, Month 2023
p-ISSN 2963-5551 http://journal.unnes.ac.id/sju/index.php/sji e-ISSN XXXX-XXXX

Prediction of Student Graduation Predicts using Hybrid 2D Convolutional Neural Network and Synthetic Minority Over-Sampling Technique

David Leandro Wibisono[1], Zaenal Abidin[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. With the rapid growth of technology, educational institutions are constantly looking for ways to improve their services and enhance student performance. One of the significant challenges in higher education is predicting the graduation outcome of students. Predicting student graduation can help educators and academic advisors to provide early intervention and support to students who may be at risk of not graduating on time. In this paper, we propose a hybrid 2D convolutional neural network (CNN) and synthetic minority over-sampling technique (SMOTE) to predict the graduation outcome of students.
Purpose: Knowing the results and how the Hybrid 2D Convolutional Neural Network (CNN) and Synthetic Minority Over-sampling Technique (SMOTE) algorithms work in predicting student graduation predicates. This algorithm uses a dataset based on family background variables and academic data.
Methods/Study design/approach: This study uses the Hybrid 2D CNN algorithm for the classification process and SMOTE for the minority class over-sampling.
Result/Findings: The prediction accuracy of the model using SMOTE is 96.31%. Meanwhile, the model that does not use SMOTE obtains an accuracy of 95.32%.
Novelty/Originality/Value: This research shows that the use of a Hybrid 2D CNN algorithm with SMOTE gives better accuracy than without using SMOTE. The dataset used also proves that family background and student academic data can be used as a reference for predicting student graduation predicates.

Keywords: Prediction, Graduation, Hybrid 2D CNN, SMOTE
Received January 27, 2023 / **Revised** February 16, 2023 / **Accepted** February 16, 2023

INTRODUCTION

Higher education has significant changes in the era of disruption, one of which is the trend of using technology in the teaching and learning process in the classroom. These changes have an impact on the success of students in receiving education. One indicator of educational success is shown from the learning achievements obtained by students, namely in the form of Grade Point Average (GPA) or *Indeks Prestasi Kumulatif* (IPK) scores when students finish studying at university [1]. One of the influences on student learning success can be sin from the family background [2]. Students who can obtain a good quality education are of course driven by their socioeconomic status. For example, such as income, education, occupation, type of residence, and ownership of property.

The relationship between family background with student graduation predicate shows that families need to preparahaood education for their children. Besides that, the government's role is also essential in helping to achieve a prosperous family. Including providing guarantees so that their children get good educational facilities. Universities also need to provide facilitation to students who are indicated at the outset to have the potential to get a poor graduation predicate.

Tens of thousands of student data were successfully recorded by Universitas Negeri Semarang (UNNES), and this data can be used for public purposes. Utilization is expected to provide useful information for universities and the community. One way to utilize this data is to process it using deep learning methods. One of the algorithms included in the deep learning method is CNN. CNN shows good performance

*Corresponding author.
Email addresses: dlwibisono.dl@students.unnes.ac.id (Wibisono)
DOI: 10.15294/rji.v1i1.65646

compared to other approaches because of its structure. The CNN structure has many different filters that can perform iterative calculations on a given input. CNN learns by making more abstract representations of data as the network structure develops deeper. CNN has been further developed by [3] with the Hybrid 2D CNN model. This algorithm provides the highest accuracy results among other algorithms such as KNN, DT, NB, Linear Regression (LR), and Artificial Neural Networks (ANN) [3]. In the deep learning method, data balance is an important factor that needs to be considered [4]. Balanced data will have a good effect on the results of the accuracy of the prediction algorithm used. While the unbalanced data itself refers to the fact that the data in each dataset class has a much different amount. To overcome the problem of data imbalance, data resampling is one method that can be used [5].

Data resampling can be done using the SMOTE algorithm. Both algorithms are two classic methods in the data resampling algorithm. SMOTE synthesizes data in the minority class with linear interpolation to increase the amount of data in the minority class [5]. In a study on system attack detection using the Network Security Layer-Knowledge Discovery in Database (NSL-KDD) dataset, good results were shown when the dataset class was balanced with the SMOTE algorithm. The results of this study were compared between the original and balanced datasets. The results of this study indicate that the accuracy of each class prediction increases in a balanced dataset [6].

Given the problem of predicting graduation predicates and the importance of a balanced dataset during the classification process, researchers will conduct research on predicting student graduation predicates using the Hybrid 2D CNN and SMOTE algorithms. The case study that will be taken in this research is the use of family background variables and academic data of UNNES students entering the 2013-2018 class who have graduated. The dataset was taken through the data request process in the UNNES Data system (https://data.unnes.ac.id).

METHODS

As research conducted by [7], datasets are important in the field of deep learning. Therefore, before the classification process, it is necessary to carry out data preprocessing to clean and balance the dataset. After that, the classification process is carried out, and the model that is successfully created will be tested using the evaluation method. The following are the steps in the method used.

Data Preprocessing

Data preprocessing is needed so that the dataset used in the study avoids inconsistent data, errors, missing values, and unbalanced. Because these things can reduce performance during the classification process. The stages carried out in data processing are shown in Figure 1.

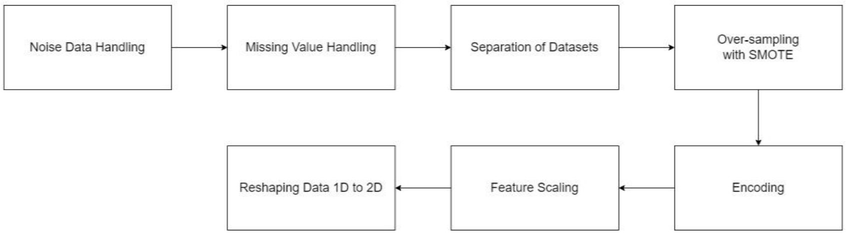


Figure 1. Stages of data preprocessing

a. Noise Data Handling
Many approaches are available to overcome problems caused by missing values in data preprocessing [8]. The first option usually discards data that may contain missing values. However, this approach is rarely profitable, because omitting data can introduce biases into the learning process, and important information can be omitted. Data imputation comes from the science of statistics. Data imputation models the probability function of the data and considers the mechanisms that cause data loss. This procedure takes approximate data from a probabilistic model to fill in the missing values.

b. Missing Value Handling
To deal with the noise problem in data mining, two main approaches are usually used in the data preprocessing literature. The first is correcting the noise using data polishing methods, especially if it

affects the labeling of data. Both use noise filters, which identify and remove data that is noisy in the training data and do not require data mining techniques to do so [9].

c. Separation of Datasets
Before classifying, the dataset needs to be separated into several parts, namely the training dataset, validation dataset, and testing dataset. The training dataset is used to train and budget-learning ing models. Then the model that has been created will be optimized using the validation dataset, the model will be tested in each training process. After the training process is complete, the model will be tested using a testing dataset. The results of testing this dataset can show the results of the accuracy of the model that has been made. The dataset distribution ratio that will be used is the training dataset of 70%, the validation dataset of 15%, and the testing dataset of 15%.

d. Over-sampling with SMOTE
According to [10], SMOTE over-sampling the minority class by taking each minority class data and including examples of artificial (synthetic) data. Synthetic data is entered along a line segment connecting each/all of the minority class's nearest neighbors. Depending on the amount of over-sampling required, synthetic data from the k-nearest neighbors are randomly selected. According to the simple explanation of [10], for example, the value of k in the specified nearest neighbors is five and the required amount of over-sampling is 200%. Then only two of the five nearest neighbors are selected. Then the resulting synthetic data is taken from each direction. Synthetic data is generated in the following way.

1. Take the difference between the vector of the data variable under consideration and its nearest neighbor.
2. Multiply this difference by a random number between 0 and 1.
3. Then add to the vector the variable under consideration.

e. Encoding
The encoding process in this study uses the one-hot encoding method. This method is used to solve problems with categorical and non-ordinal type variables. Each variable will be broken down into several new variables according to the categories in the variable, then the value will be in binary form. It can also increase the number of variables used in the classification process so that data can be represented as a two-dimensional image. While the labels on the data are of category type and ordinal form. Due to the small number of class labels, a simple encoding process can be carried out by changing the data category to a certain number in order from the lowest to the highest pass predicate.

f. Feature Scaling
Feature scaling or standardization is one of the stages in data preprocessing which normalizes data by converting it into a certain range. Normalized data can accelerate calculations in the classification algorithm [11]. Feature scaling on data is generally done using Keras, Scikit-learn, and deep learning. One of the commonly used feature scaling techniques is MinMaxScaler. This technique scales the values in the data into the range of 0 to 1. The equation for the MinMaxScaler process is shown in Equation 1.

$$MinMaxScaler = \frac{X - Xmin}{Xmax - Xmin} \tag{1}$$

X is a representation of the original value, $Xmin$ is the smallest value of all data in a variable, and $Xmax$ is the largest value of all data in a variable. MinMaxScaler will be used on variables with ordinal values, namely numbers that can be sorted from largest to smallest value.

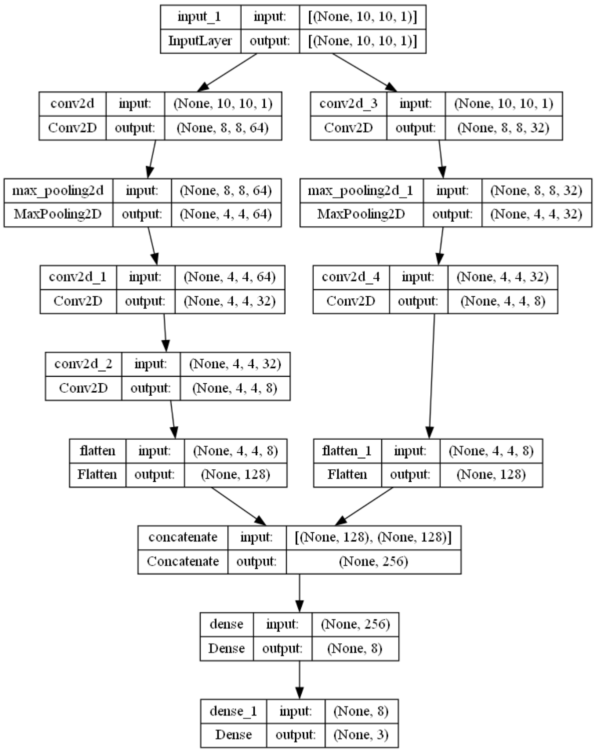
g. Reshaping Data 1D to 2D
In this study, the Hybrid 2D CNN algorithm requires 2D data, but the data obtained from the data source is 1D. Therefore, the data needs to be converted into 2D form. This process can be done using the reshape function from the NumPy library.

Classification

a. Determination of Training Parameters
Before classifying, it is necessary to determine some training parameters that will be used. Some of the training parameters determined in this study are learning rate, optimizer, and epoch. The learning rate used is 0.001, the Optimizer type is Adaptive Moment (Adam), and the epoch is 100.

b. Classification Process

Hybrid 2D CNN applies two different 2D CNN models to the same classification process. The number of hidden layers used between the two algorithms can be different or the same. This research uses a model that has been developed by [3], the Hybrid 2D CNN model used is shown in Figure 2.



input_1 | input: | [(None, 10, 10, 1)]
InputLayer | output: | [(None, 10, 10, 1)]
conv2d | input: | (None, 10, 10, 1)
Conv2D | output: | (None, 8, 8, 64)
conv2d_3 | input: | (None, 10, 10, 1)
Conv2D | output: | (None, 8, 8, 32)
max_pooling2d | input: | (None, 8, 8, 64)
MaxPooling2D | output: | (None, 4, 4, 64)
max_pooling2d_1 | input: | (None, 8, 8, 32)
MaxPooling2D | output: | (None, 4, 4, 32)
conv2d_1 | input: | (None, 4, 4, 64)
Conv2D | output: | (None, 4, 4, 32)
conv2d_4 | input: | (None, 4, 4, 32)
Conv2D | output: | (None, 4, 4, 8)
conv2d_2 | input: | (None, 4, 4, 32)
Conv2D | output: | (None, 4, 4, 8)
flatten | input: | (None, 4, 4, 8)
Flatten | output: | (None, 128)
flatten_1 | input: | (None, 4, 4, 8)
Flatten | output: | (None, 128)
concatenate | input: | [(None, 128), (None, 128)]
Concatenate | output: | (None, 256)
dense | input: | (None, 256)
Dense | output: | (None, 8)
dense_1 | input: | (None, 8)
Dense | output: | (None, 3)

Figure 2. Hybrid 2D CNN model [3]

In Figure 2, the model starts with the input layers. Then the process is divided into two different 2D CNN models. The left side has five layers and the right side has four layers. The various layers in the two models are convolution layers and pooling layers. At the end of the model, the two outputs are flattened to become 1D data. After that, it is put together and processed on fully connected layers. Next will be explained in more detail about the CNN process on each layer.

Evaluation

The classification that has been done, will produce a model. This model is used for the process of predicting student graduation predicates. In practice, the model cannot be immediately considered successful, if it has not been evaluated. The evaluation method that will be used in this study is the confusion matrix.

RESULT AND DISCUSSION

The deep learning model used was built using two different CNN models. The combination of the two models is referred to as Hybrid 2D CNN [3]. In the dataset, there are variables and labels. The variables used consist of 31 variables. While the labels in the dataset consist of three classes. The three classes are student graduation and predicate classes. The graduation predicate consists of with honors, very satisfying, and satisfying. The description and number of each class on the label can be seen in Table 1.

Table 1. Label description on dataset

No	Class Name	Description	Total
1	With Praise	The graduation predicate is obtained by students with a GPA of more than 3.50 and passing less than or equal to nine semesters.	11.269
2	Very Satisfied	Graduate predicates are obtained by students with a GPA of more than 2.75 and less than or equal to 3.50. Or students who graduate with a GPA of more than 3.50 but have a length of study of more than nine semesters.	13.087
3	Satisfied	Graduate predicates are obtained by students with a GPA of more than or equal to 2.00 and less than or equal to 2.75.	27
		Grand Total	24.383

The variable itself consists of three categories, namely student background, family background, and academic data. The student background variable used in this study was only student gender because this variable has an important consideration in determining student academic achievement [12]. More details regarding the description of the variables in the dataset used are shown in Table 2.

Table 2. Description of variables in the dataset

No	Variable Name	Description	Category	Type
1	jenis_kelamin	Student Gender	Student background	*Binary*
2	ayah_hidup	The condition of student's father is still alive	Family's background	*Binary*
3	ibu_hidup	The condition of student's mother is still alive	Family's background	*Binary*
4	fakultas	The name of the faculty where the student is studying	Family's background	*Nominal*
5	pekerjaan_ayah	The job owned by the student's father	Family's background	*Nominal*
6	pekerjaan_ibu	The job owned by the student's mother	Family's background	*Nominal*
7	pendidikan_ayah	The last education was taken by the student's father	Family's background	*Nominal*
8	pendidikan_ibu	The last education was taken by the student's mother	Family's background	*Nominal*
9	ukt	The amount of UKT that must be paid by students every semester to the university	Family's background	*Numeric*
10	urutan_kelahiran	Order of birth from students to siblings	Family's background	*Nominal*
11	beasiswa	Status whether the student has a scholarship or not	Family's background	*Numeric*
12	jml_mk_mengulang	The number of courses repeated by students during college	Academic data	*Numeric*
13	jml_sks_mengulang	The number of credits that are repeated by students during college	Academic data	*Numeric*
14	jml_cuti	The number of studies leaves that students take	Academic data	*Numeric*
15	jml_registrasi	The number of registrations carried out by students is counted once every semester	Academic data	*Numeric*
16	jml_tidak_registrasi	The number of unregistered students is counted once every time they leave one semester before graduating	Academic data	*Numeric*
17	ips_1	IPS students in semester 1	Academic data	*Numeric*
18	ips_2	IPS students in semester 2	Academic data	*Numeric*
19	ips_3	IPS students in semester 3	Academic data	*Numeric*
20	ips_4	IPS students in semester 4	Academic data	*Numeric*
21	ips_5	IPS students in semester 5	Academic data	*Numeric*
22	ips_6	IPS students in semester 6	Academic data	*Numeric*
23	ips_7	IPS students in semester 7	Academic data	*Numeric*
24	ips_8	IPS students in semester 8	Academic data	*Numeric*
25	ips_9	IPS students in semester 9	Academic data	*Numeric*
26	ips_10	IPS students in semester 10	Academic data	*Numeric*
27	ips_11	IPS students in semester 11	Academic data	*Numeric*
28	ips_12	IPS students in semester 12	Academic data	*Numeric*
29	ips_13	IPS students in semester 13	Academic data	*Numeric*
30	ips_14	IPS students in semester 14	Academic data	*Numeric*
31	lama_studi	Length of study in semesters taken by students	Academic data	*Numeric*

The 24,383 datasets will be divided into three parts, namely 70% for the training dataset, 15% for the validation dataset, and 15% for the testing dataset. This is so that deep learning models can learn using training data and are not contaminated with validation and testing data during training. The amount of data for each class that has been divided is shown in Table 3.

Table 3. Total data for each class

No	Data Type	With Praise	Very Satisfied	Satisfied
1	Total Training Data	7.888	9.161	19
2	Total Validation Data	1.690	1.963	4
3	Total Testing Data	1.691	1.963	4
	Grand Total	11.269	13.087	27

After the dataset is divided, the index of each type of data will be stored. Next is the process of over-sampling the minority class using SMOTE. This process uses the entire initial dataset before the division is carried out. A comparison of the amount of data before and after SMOTE is shown in Table 4.

Table 4. Total data for each class before and after SMOTE

No	Dataset	With Praise	Very Satisfied	Satisfied
1	Before SMOTE	11.269	13.087	27
2	After SMOTE	11.269	13.087	13.087

After SMOTE, the process of encoding and feature scaling is carried out. Then the dataset results of the process will be sorted. Next, the ENN process is carried out on the minority class training dataset. The final amount can be seen in Table 5.

Table 5. Total minority class before SMOTE, after SMOTE, and after Separated

No	Description	Total
1	Before SMOTE	19
2	After SMOTE	13.087
3	After being separated from validation and testing datasets	13.079

The final step is to reshape the data into 2D, and the training process is carried out. The training process can be seen in Figure 3. This process was carried out in 100 epochs. The results shown in Figure 3 shows that the level of accuracy of the training and validation processes does not have a big difference. So it can be said that the results of this training are not overfitting.

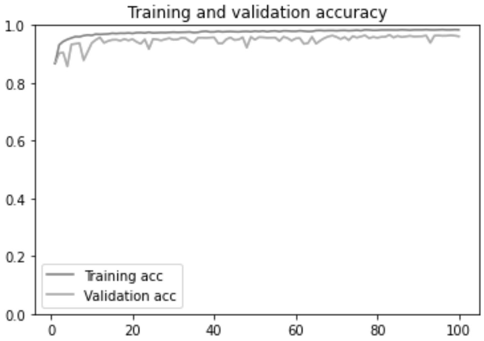


Figure 3. InfoGraphic classification process accuracy

After that, an evaluation process is carried out using the validation dataset. The evaluation results will then be represented in the form of a Confusion Matrix which can be seen in Figure 4. The labels in Figure 4 can be explained as follows: M is Satisfactory, SM is Very Satisfying, and DP is With Praise. The Satisfying class was successfully predicted correctly in four data out of the four existing data. The Very Satisfying Class was successfully predicted correctly in 1888 of the 1963 data available. The Class with Praise was successfully predicted correctly in as many as 1624 of the 1691 data.

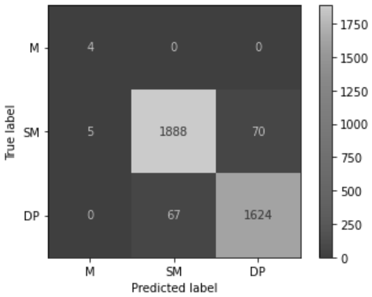


Figure 4. Confusion matrix evaluation results

The research conducted uses a dataset without the SMOTE process and a dataset with SMOTE. From the evaluation results carried out on the two different datasets, it was found that the use of the data balancing method using SMOTE succeeded in increasing the accuracy of predicting student graduation predicates. More details can be seen in Table 6.

Table 6. Comparison of accuracy based on SMOTE usage

No	SMOTE Use	Accuracy
1	Without SMOTE	95,32%
2	With SMOTE	96,31%

CONCLUSION

Based on the results of research that has been conducted regarding the predicate of student graduation using a Hybrid 2D Convolutional Neural Network and Synthetic Minority Over-Sampling Technique, the following conclusions can be drawn.

1) Implementation of the 2D CNN Hybrid Algorithm and SMOTE to predict the predicate of student graduation based on family background and academic data is carried out with several stages. The stages begin with the data collection process, preprocessing data, data classification, and evaluation. The process of preprocessing data respectively consists of handling noise data, handling missing values, dataset separation, over-sampling with SMOTE, encoding data, feature scaling with MinMaxScaler, and reshaping data into 2D. Whereas the data classification process is carried out with the stages of determining the training parameters, making 2D CNN hybrid models, and the training process. Finally, the evaluation process is carried out by the confusion matrix method. Some of the things used in this study are the ratio of dataset distribution of 70% for dataset training, 15% for dataset validation, and 15% for dataset testing, the use of learning rates of 0.001, epoch amounting to 100, adam optimizer, and the use of the balanced dataset with a SMOTE algorithm.
2) Predicate Student graduation based on family background and academic data using the 2D CNN hybrid algorithm and SMOTE produces an accuracy of 96.53%. These results are better than the training process on a dataset that are not balanced using SMOTE. The model that uses dataset with the application of SMOTE has an accuracy of 96.31%. While the model that uses dataset without the application of SMOTE has an accuracy of 95.32%.

In future research, an algorithm can be used to under-sampling synthetic data on minority classes that have been oversampled. So that it can be seen whether the under-sampling algorithm can improve the accuracy of the model that has been made. “Analisis Fakto-faktor yang Mempengaruhi Prestasi Belajar Pada Mahasiswa Semester II Program Studi DIII Kebidanan Stikes ‘Aisyiyah Yogyakarta Tahun 2013 1,” 2013.

REFERENCES

[1] R. K. Niswatin, R. Wulanningrum, and U. Syaidah, “Sistem Prediksi Nilai IPK Mahasiswa Menggunakan Metode K-Nearest Neighbor,” *Jurnal Maklumatika*, vol. 3, no. 1, 2016.

[2] F. Kusumawati and E. Nurhidayati, “Analisis Faktor-faktor yang Mempengaruhi Prestasi Belajar Pada Mahasiswa Semester II Program Studi DIII Kebidanan Stikes ‘Aisyiyah Yogyakarta Tahun 2013 1,” 2013.

[3] S. Poudyal, M. J. Mohammadi-Aragh, and J. E. Ball, “Prediction of Student Academic Performance Using a Hybrid 2D CNN Model,” *Electronics (Switzerland)*, vol. 11, no. 7, Apr. 2022, doi: 10.3390/electronics11071005.

[4] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, “Learning from class-imbalanced data: Review of methods and applications,” *Expert Systems with Applications*, vol. 73. Elsevier Ltd, pp. 220–239, May 01, 2017. doi: 10.1016/j.eswa.2016.12.035.

[5] Z. Xu, D. Shen, T. Nie, and Y. Kou, “A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data,” *J Biomed Inform*, vol. 107, Jul. 2020, doi: 10.1016/j.jbi.2020.103465.

[6] X. Zhang, J. Ran, and J. Mi, “An Intrusion Detection System Based on Convolutional Neural Network for Imbalanced Network Traffic,” in *IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, 2019, pp. 456–460.

[7] S. Krishnan, M. J. Franklin, K. Goldberg, J. Wang, and E. Wu, “ActiveClean: An interactive data cleaning framework for modern machine learning,” in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Jun. 2016, vol. 26-June-2016, pp. 2117–2120. doi: 10.1145/2882903.2899409.

[8] J. Luengo, S. García, and F. Herrera, “On the choice of the best imputation methods for missing values considering three groups of classification methods,” *Knowl Inf Syst*, vol. 32, no. 1, pp. 77–108, Jul. 2012, doi: 10.1007/s10115-011-0424-2.

[9] B. Frénay and M. Verleysen, “Classification in the presence of label noise: A survey,” *IEEE Trans Neural Netw Learn Syst*, vol. 25, no. 5, pp. 845–869, 2014, doi: 10.1109/TNNLS.2013.2292894.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[11] T. D. K. Thara, P. S. Prema, and F. Xiong, “Auto-detection of epileptic seizure events using deep neural network with different feature scaling techniques,” *Pattern Recognit Lett*, vol. 128, pp. 544–550, Dec. 2019, doi: 10.1016/j.patrec.2019.10.029.

[12] C.-C. Kiu, “Data Mining Analysis on Student’s Academic Performance through Exploration of Student’s Background and Social Activities,” in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, 2018.