

Classification of deep learning convolutional neural network feature extraction for student graduation prediction

Abu Salam, Junta Zeniarja

Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

Article Info

Article history:

Received Nov 29, 2022

Revised Jun 7, 2023

Accepted Jun 17, 2023

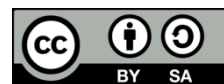
Keywords:

Classification
Convolutional neural network
Deep learning
Graduation
Student

ABSTRACT

One indicator of a university's educational quality is the proportion of enrolled students who actually graduate within four years. This proportion is typically fewer than the number of students that enroll in a given year. A low graduation rate can have a negative impact on both the university's reputation and its accreditation because it indicates that fewer students are completing their degrees. Student activity, economic, and other issues all play a role in why some students are unable to complete their degrees on time. As a result, stakeholders need a model that can predict whether or not students will graduate on time as a means of evaluating and giving a basis for policy actions. This research proposes a model for converting textual data into an image format using a deep learning convolutional neural network (CNN), and then classifying the extracted features using a variety of machine learning classification algorithms like the decision tree, random forest, Naive Bayes, support vector machine (SVM), and k-nearest neighbor (K-NN). The classification model trained on feature extraction data had a 96.1% accuracy rate, while the classification model trained on the original data achieved a 71.2% accuracy rate.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Abu Salam

Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro
Semarang, Indonesia

Email: abu.salam@dsn.dinus.ac.id

1. INTRODUCTION

Since education plays an inextricable part in a country's development, universities work hard to provide top-notch instruction that will benefit students in the long run [1]–[3]. The primary mission of universities is to produce highly skilled workers who can compete successfully in the labor market; graduates can then use these acquired abilities to secure employment and advance their careers [4]. As a result, many students run out of time or simply don't manage to do their coursework at all [5]. Despite the fact that students' academic achievements matter greatly, universities have other concerns as well. This is due to the fact that a high rate of student success in the classroom is a hallmark of a high-quality university. Increasing the number of students who do not graduate on time can lead to an increase in the amount of academic data from all students who are still enrolled, potentially threatening the reputation and accreditation value of the tertiary institution. Authorities in postsecondary institutions need to predict when students will graduate as a control and anticipation measure, and the process of doing so is also a step toward anticipating dropout problems, which are a serious issue in an education system because they cause financial, social, and economic losses. Students and the government are the primary actors in this scenario, thus we will focus on the political, academic, and economic dimensions of education [6]. In order to gain insight from data given as information, data mining (DM) has been shown to be useful [7]. The scope and variety of schooling data make it challenging to process

accurately. Data preparation is a crucial part of the educational data mining (EDM) process that can yield the best results and EDM can be used as a solution to this issue [8], [9]. Neural networks [10], Naive Bayes, decision trees, k-nearest neighbor (K-NN), support vector machine (SVM), and discriminant analysis are all examples of often-used classification methods in EDM research [11]–[15]. Predicting student success in higher education using machine learning models has become common practice. Attribute data collected through academic processes and saved in custom-built applications or databases are processed using a variety of machine learning (ML) algorithms [7], [16].

Multi-layered artificial neural networks like the deep learning convolutional neural network (CNN) model are frequently employed in image processing [17]. The CNN is a popular deep learning model due to its high performance in a variety of use cases. The field of education is one that has recently seen an uptick in interest in deep learning [17], [18]. While CNN has been widely hailed for its ability to analyze and recognize images, it has also been shown that the feature extraction process using CNN accurately represents the original data form in many other contexts, including educational data mining. Each neuron in a CNN receives multiple inputs, generates a product point value, applies an activation function, and finally, in the final (fully connected) layer, there is a loss function that measures the discrepancy between the predicted value and the expected output value [17], [19]. The hypothesis that the image data format allows for a better representation of the relationships between features that can be recognized by the CNN for analysis and prediction processes provides support for the process of transforming non-image data into the image data format [19]. In order to better express the relationship between features, such as categories or feature similarities, the sequence of features is sometimes reorganized in 2-D space during the transformation of tabular (non-image) data [20], [21]. Tabular data can be used in the feature extraction process with deep learning convolutional neural networks to predict graduation, and it has been shown to produce better results than conventional data mining, with the best accuracy of 77.35% using the random forest algorithm, and with deep learning CNN for the same data to achieve the achievement value of 87.44%.

In this paper, we propose a model for converting tabular data into image format using a deep learning CNN, and we'll classify the extracted features using a variety of machine learning algorithms like the decision tree, random forest, Naive Bayes, SVM, and K-NN. The data used is comprised of 4,041 grads from 4 different programs at the bachelor's level from the Faculty of Computer Science at Universitas Dian Nuswantoro. The best values of recall, precision, and f-measures for classification will be determined by comparing the machine learning model's output on the original data with that obtained using the CNN deep learning feature.

2. METHOD

2.1. Dataset

There were a total of 4,041 records used in the research datasets, all of which related to graduates of undergraduate study programs at the Faculty of Computer Science, Universitas Dian Nuswantoro, with data collected from 2012-2018 for a total of 4 different programs. There are 2,293 records for study code 11, 658 for study code 12, 666 for study code 14, and 424 for study code 15. A comparison graph of the number of labels included in the dataset is shown in Figure 1, which also has approximately the same amount of data. Label 1 has the equivalent of roughly 2,073 records (covering a research duration of less than 48 months and no more than 4 years). Label 2 contains 1,968 records (studied for more than 48 months; delayed by more than 4 years).

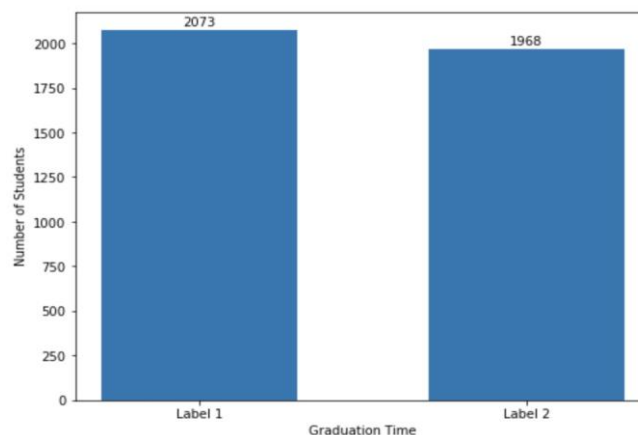


Figure 1. The number of records in the dataset

Details for each of the descriptions of the attributes that were used can be found in Table 1, which contains information about the dataset attributes, which include age, marital status, place of birth, number of scholarships, number of applications for leave, number of student activities, number of achievements, GPA1, GPA2, GPA3 and GPA4. In addition, Table 1 contains information about the number of achievements that were earned by each student.

Table 1. Attribute of dataset

Attribute name	Description
Study program	4 undergraduate study programs (A11, A12, A14, and A15)
Hometown	1: for domicile in the city 2: for domicile outside the city
Age	1: age >0 and age <12 2: age >=12 and age <26 3: age >=26 and age <46 4: age >= 6
Marital	1: Married 2: Not married
Number of scholarships	Number of scholarships received
Number of leave applications	Number of times you have applied for leave
Number of student activities	Number of student activities participated in
Number of achievements	Number of achievements or certificates ever obtained
GPA1	Grade 1 for grade point average
GPA2	Grade 2 for grade point average
GPA3	Grade 3 for grade point average
GPA4	Grade 4 for grade point average

2.2. Research step

Only numerical features can be processed by deep learning models [22]. One type of deep learning model used in the training and testing of data assigned a label is CNN [23], while the goal of the machine learning model is to achieve the best possible accuracy through experimentation [24]. In order to obtain the best accuracy value from the newly created features during training, this study will employ the deep learning model for feature extraction during the data training process, followed by the machine learning model for the classification process. Figure 2 explains the main steps in the research phase carried out.

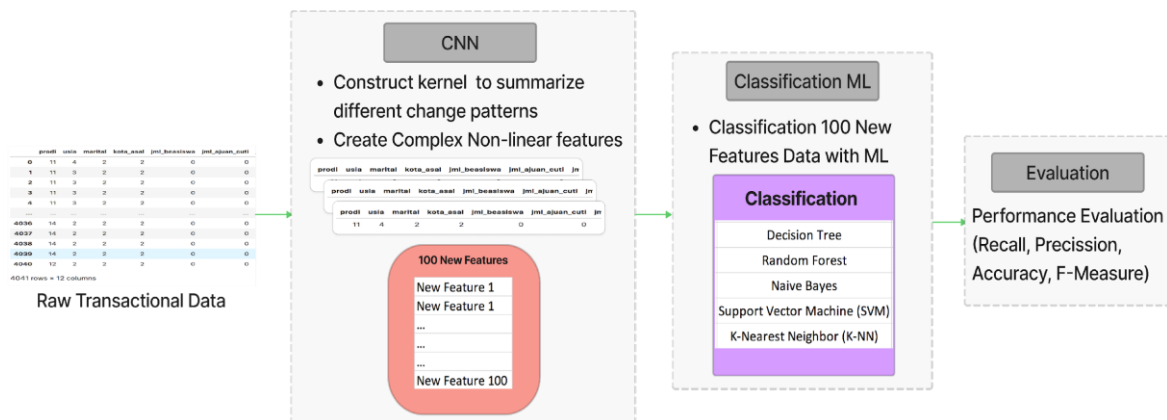


Figure 2. Method steps

In the proposed research methodology, out of 4,041 dataset records consisting of 12 attributes to be processed with the CNN architecture, the attribute grouping is carried out as follows, from N is the number of data records as many as 4,041, a is an attribute that is grouped into 3 categories as follows:

- Student personal data attributes: (age, marital, city of origin, and scholarship).
- Historical attributes: (number of leave requests, amount of arrears, number of student activities, and number of achievements/certificate awards).
- Academic value attributes: (semester 1-semester 4 achievement index: IPS1, IPS2, IPS3, IPS4).

And W is the number of attributes from each group from A and K is worth 1 for the total category of graduation data from the Faculty of Computer Science at Universitas Dian Nuswantoro. So that the input data for the CNN deep learning stage is $N \times A \times W \times K = 4,041 \times 3 \times 4 \times 1$. Additionally, classification ML will be implemented using 5 classification techniques, including decision tree, random forest, Naive Bayes, SVM, and K-NN, to obtain the best evaluation value from the 100 new features data generated. Measures of recall, precision, accuracy, and f-measure are utilized in the evaluation process [25], [26].

3. RESULTS AND DISCUSSION

Figure 3 depicts the fundamental components of a CNN, including the convolution 2D layer, the 2D pooling layer, the batch normalization layer, the fully connected layer, the non-linear activation function ReLU, and the max pooling 2D layer. In the following example, illustrated in Figure 4, we see the result of feature extraction data collected from the CNN stages for feature extraction on dense (dense feature) with output shape (none, 100). Table 2 displays the results of applying the machine learning classification method to the data resulting from feature extraction.

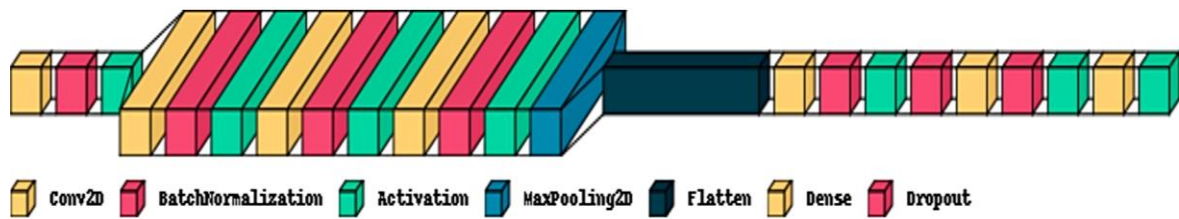


Figure 3. CNN stages

9	...	92	93	94	95	96	97	98	99
0.516714	...	0.300420	-1.539349	2.059091	-0.181639	1.774889	-0.294648	-0.005839	-1.971039
-0.429307	...	0.047781	-0.437921	1.042610	-0.294263	1.459267	-0.145321	0.181368	-0.677201
-0.352103	...	-0.113610	-0.476321	1.110847	-0.764649	1.287692	-0.077738	-0.524138	-0.677101
-0.681955	...	-0.399359	-0.635935	2.034534	0.153822	-0.351255	0.034884	0.330606	-1.114779
0.325843	...	0.277672	-0.399737	0.637576	-0.416581	1.033822	-0.914346	0.160373	-0.857471
1.273204	...	0.076275	-0.892636	0.293125	-0.671600	-0.016163	-0.912109	-0.090875	-1.692824
0.062645	...	0.797004	-0.639692	0.619118	-0.343200	1.459566	-1.052675	0.151528	-0.614814
1.229106	...	-0.106534	-0.354537	0.545832	-0.037908	0.067480	-1.100398	0.453499	-1.278038
0.272940	...	0.087951	-0.705217	1.441743	0.469144	-0.243263	-0.574374	0.441524	-1.461822
-0.741434	...	-0.014953	-0.223610	1.711095	0.290198	0.125478	-0.256920	0.205205	-1.292538

Figure 4. Sample data for 100 new features

Table 2. Classification results

No	Model	Accuracy	Recall	Precision	F1-score
1	Naive Bayes CNN features	0.961253	0.961253	0.963212	0.961237
2	Naive Bayes actual features	0.708986	0.708986	0.713551	0.708235
3	SVM linear CNN features	0.956307	0.956307	0.957542	0.956299
4	SVM linear actual features	0.710635	0.710635	0.715877	0.709730
5	Decision tree CNN features	0.954658	0.954658	0.958495	0.954604
6	Decision tree actual features	0.723825	0.723825	0.727725	0.721801
7	KNN linear CNN features	0.938170	0.938170	0.938547	0.938174
8	KNN linear actual features	0.706513	0.706513	0.706571	0.706534
9	Random forest CNN features	0.938170	0.938170	0.938547	0.938174
10	Random forest actual features	0.706513	0.706513	0.706571	0.706534

Table 2 shows that the data classification process based on the CNN feature extraction significantly improved over the classification process based on actual features in terms of accuracy, recall, precision, and f1-score. Figure 5 displays the accuracy performance rating. While the KNN algorithm achieves the highest accuracy for actual features 74.8%, the SVM algorithm achieves the highest accuracy for training data 95.1%.

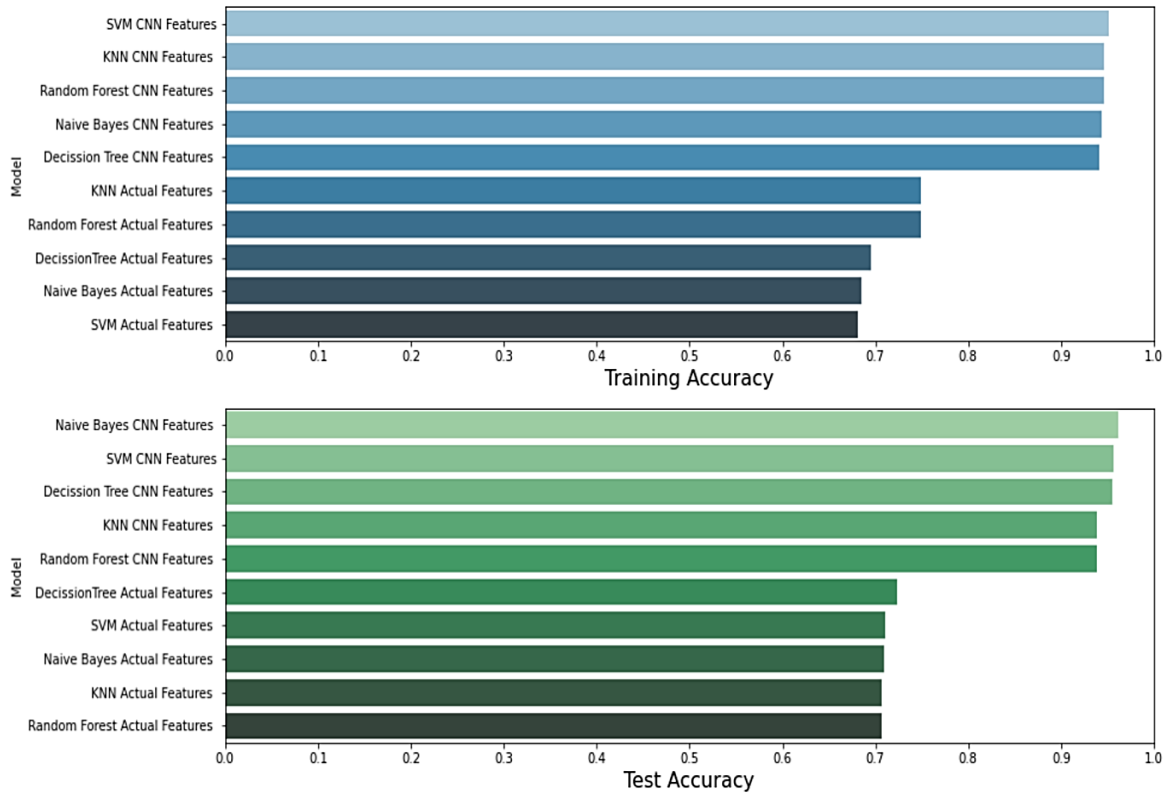


Figure 5. Training and testing accuracy

Based on Table 3, the Naive Bayes algorithm achieves the best data testing accuracy value for the CNN feature of 96.1%, compared to only using actual features which achieves an accuracy value of 70.9%. In addition, the SVM algorithm also achieves the best data testing accuracy value for the CNN feature 95.6% compared to only using actual features which is 71%. Then the decision tree algorithm also achieves an accuracy value of data testing for CNN features of 95.5%, compared to only using actual features which is 72.4%. Likewise with the use of the KNN and random forest algorithms which have higher accuracy when using CNN features than those that only use actual features. So the use of the CNN feature plays a very large role in increasing accuracy compared to those that only use actual features.

Table 3. Comparison of accuracy values

No	Model	Training accuracy	Testing accuracy
1	Naive Bayes CNN features	0.944130	0.961253
2	Naive Bayes actual features	0.685290	0.708986
3	SVM CNN features	0.951556	0.956307
4	SVM actual features	0.681400	0.710635
5	Decision tree CNN features	0.940948	0.954658
6	Decision tree actual features	0.695191	0.723825
7	KNN CNN features	0.945898	0.938170
8	KNN actual features	0.748586	0.706513
9	Random forest CNN features	0.945898	0.938170
10	Random forest actual features	0.748586	0.706513

4. CONCLUSION

In this paper, we explain the process of transforming data from its raw form into a data form that still accurately depicts the raw data. It has been demonstrated that implementing a deep learning CNN for the process of feature extraction with the CNN will boost the accuracy of the classification results achieved by all machine learning algorithms that have been implemented. In the future, the authors will undertake experiments for data complexity with a greater and more diversified number of characteristics and record power. They will also be able to compare the feature selection process for the attributes that will be employed, as well as the hyperparameters for the CNN architecture that was used.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Universitas Dian Nuswantoro and Dinustek for providing us with a space in which we are able to conduct this research in an appropriate manner.




REFERENCES

- [1] A. J. Fernández-García, R. Rodríguez-Echeverría, J. C. Preciado, J. M. C. Manzano, and F. Sánchez-Figueroa, "Creating a recommender system to support higher education students in the subject enrollment decision," *IEEE Access*, vol. 8, pp. 189069–189088, 2020, doi: 10.1109/ACCESS.2020.3031572.
- [2] A. Nabil, M. Seyam, and A. Abou-Elfetouh, "Prediction of students' academic performance based on courses' grades using deep neural networks," *IEEE Access*, vol. 9, pp. 140731–140746, 2021, doi: 10.1109/ACCESS.2021.3119596.
- [3] C. Jin, B. Li, S. J. T. Jansen, H. J. F. M. Boumeester, and P. J. Boelhouwer, "What attracts young talents? Understanding the migration intention of university students to first-tier cities in China," *Cities*, vol. 128, p. 103802, Sep. 2022, doi: 10.1016/j.cities.2022.103802.
- [4] A. Salam, J. Zeniarja, and D. M. Anthareza, "Student graduation prediction model using deep learning convolutional neural network (CNN)," in *2022 International Seminar on Application for Technology of Information and Communication: Technology 4.0 for Smart Ecosystem: A New Way of Doing Digital Business, iSemantic 2022*, 2022, pp. 362–366, doi: 10.1109/iSemantic55962.2022.9920449.
- [5] S. Syahrudin *et al.*, "Students' acceptance to distance learning during Covid-19: the role of geographical areas among Indonesian sports science students," *Heliyon*, vol. 7, no. 9, p. e08043, Sep. 2021, doi: 10.1016/j.heliyon.2021.e08043.
- [6] J. Niyogisubizo, L. Liao, E. Nziyumva, E. Murwanashyaka, and P. C. Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100066, 2022, doi: 10.1016/j.caeai.2022.100066.
- [7] A. Yunita, H. B. Santoso, and Z. A. Hasibuan, "Everything is data: towards one big data ecosystem using multiple sources of data on higher education in Indonesia," *Journal of Big Data*, vol. 9, no. 1, p. 91, Dec. 2022, doi: 10.1186/s40537-022-00639-7.
- [8] G. Feng, M. Fan, and C. Ao, "Exploration and visualization of learning behavior patterns from the perspective of educational process mining," *IEEE Access*, vol. 10, pp. 65271–65283, 2022, doi: 10.1109/ACCESS.2022.3184111.
- [9] M. M. Rahman, Y. Watanobe, T. Matsumoto, R. U. Kiran, and K. Nakamura, "Educational data mining to support programming learning using problem-solving data," *IEEE Access*, vol. 10, pp. 26186–26202, 2022, doi: 10.1109/ACCESS.2022.3157288.
- [10] J. T. Hancock and T. M. Khoshgoftar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, no. 1, p. 28, Dec. 2020, doi: 10.1186/s40537-020-00305-w.
- [11] Y. Nieto, V. Gacia-Diaz, C. Montenegro, C. C. Gonzalez, and R. G. Crespo, "Usage of machine learning for strategic decision making at higher educational institutions," *IEEE Access*, vol. 7, pp. 75007–75017, 2019, doi: 10.1109/ACCESS.2019.2919343.
- [12] M. A. Prada *et al.*, "Educational data mining for tutoring support in higher education: a web-based tool case study in engineering degrees," *IEEE Access*, vol. 8, pp. 212818–212836, 2020, doi: 10.1109/ACCESS.2020.3040858.
- [13] K. L. M. Ang, F. L. Ge, and K. P. Seng, "Big educational data analytics: survey, architecture and challenges," *IEEE Access*, vol. 8, pp. 116392–116414, 2020, doi: 10.1109/ACCESS.2020.2994561.
- [14] T. T. Mai, M. Bezbradica, and M. Crane, "Learning behaviours data in programming education: community analysis and outcome prediction with cleaned data," *Future Generation Computer Systems*, vol. 127, pp. 42–55, Feb. 2022, doi: 10.1016/j.future.2021.08.026.
- [15] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, p. 11, Dec. 2022, doi: 10.1186/s40561-022-00192-z.
- [16] M. Akour, H. A. Sghaier, and O. A. Qasem, "The effectiveness of using deep learning algorithms in predicting students achievements," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 19, no. 1, pp. 388–394, Jul. 2020, doi: 10.11591/ijeecs.v19.i1.pp388-394.
- [17] A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, and B. Navarro-Colorado, "A systematic review of deep learning approaches to educational data mining," *Complexity*, vol. 2019, pp. 1–22, May 2019, doi: 10.1155/2019/1306039.
- [18] A. S. Aljaloud *et al.*, "A deep learning model to predict student learning outcomes in LMS using CNN and LSTM," *IEEE Access*, vol. 10, pp. 85255–85265, 2022, doi: 10.1109/ACCESS.2022.3196784.
- [19] Y. Zhu *et al.*, "Converting tabular data into images for deep learning with convolutional neural networks," *Scientific Reports*, vol. 11, no. 1, p. 11325, May 2021, doi: 10.1038/s41598-021-90923-y.
- [20] A. Sharma, E. Vans, D. Shigemizu, K. A. Boroevich, and T. Tsunoda, "DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture," *Scientific Reports*, vol. 9, no. 1, p. 11399, Aug. 2019, doi: 10.1038/s41598-019-47765-6.
- [21] O. Bazgir, R. Zhang, S. R. Dhruva, R. Rahman, S. Ghosh, and R. Pal, "Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks," *Nature Communications*, vol. 11, no. 1, p. 4391, Sep. 2020, doi: 10.1038/s41467-020-18197-y.
- [22] H. Prabowo, A. A. Hidayat, T. W. Cenggoro, R. Rahutomo, K. Purwandari, and B. Pardamean, "Aggregating time series and tabular data in deep learning model for University Students' GPA Prediction," *IEEE Access*, vol. 9, pp. 87370–87377, 2021, doi: 10.1109/ACCESS.2021.3088152.




- [23] G. Feng, M. Fan, and Y. Chen, "Analysis and prediction of students' academic performance based on educational data mining," *IEEE Access*, vol. 10, pp. 19558–19571, 2022, doi: 10.1109/ACCESS.2022.3151652.
- [24] H. E. Abdelkader, A. G. Gad, A. A. Abohany, and S. E. Sorour, "An efficient data mining technique for assessing satisfaction level with online learning for higher education students during the COVID-19," *IEEE Access*, vol. 10, pp. 6286–6303, 2022, doi: 10.1109/ACCESS.2022.3143035.
- [25] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, p. e01250, Feb. 2019, doi: 10.1016/j.heliyon.2019.e01250.
- [26] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi: 10.1109/ACCESS.2020.2981905.

BIOGRAPHIES OF AUTHORS



Abu Salam    is a permanent lecturer at the Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia. He is also the software manager in the company and produces many software products for various clients. He received his master's degree in Informatics Engineering from Dian Nuswantoro University in Indonesia. Main research interests are data mining, machine learning, deep learning, information retrieval, and software engineering. His current research covers various data mining and software engineering applications, such as student graduation prediction applications, attendance applications with deep learning, classification of government document archives, sentiment analysis, and many more. He develops software to solve problems in various government agencies and companies using various models of artificial intelligence, machine learning, and deep learning. In addition, he is also active in career development activities in building the skills of students who intern at companies. He can be contacted at email: abu.salam@dsn.dinus.ac.id.



Junta Zeniarja    is a permanent lecturer at the Department of Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia. He received two master's degrees in Informatics Engineering from Universiti Teknikal Malaysia Melaka (UTeM) in Malaysia and Universitas Dian Nuswantoro in Indonesia. The main research interests are data mining, machine learning, deep learning, information retrieval, and geographic information systems. His current research covers a wide range of data mining and data science applications, such as predicting student graduation, search engines for kids, thesis document classification, sentiment analysis, and geographic information systems based on information retrieval (geographic information retrieval). He tries to produce the best predictive model by using various machine learning and deep learning techniques with various preprocessing methods such as feature extraction, feature engineering, and feature selection. In addition to data, he also tries to integrate information from the text and spatial data, which are used to explain information searches and locations. He can be contacted at email: junta@dsn.dinus.ac.id.