Prediction of Computer Science Students' Length of Study Based on GPA and Programming Course Grades using Neural Networks Algorithm

Marvel Lim Susanto
Computer Science Department
Bina Nusantara University
Jakarta, Indonesia
marvel.susanto@binus.ac.id

Computer Science Department
Bina Nusantara University
Jakarta, Indonesia
ruben.saputra@binus.ac.id

Marvelio Chandra
Computer Science Department
Bina Nusantara University
Jakarta, Indonesia
marvelio.chandra@binus.ac.id

Ruben Benedict Saputra

***Abstract*— The duration of a student's study in computer science programs can greatly vary and is influenced by various factors. This paper focuses on predicting the length of study for computer science students based on their academic factors such as grade point average (GPA) and programming course grades. To achieve this, a neural network algorithm is employed to develop a predictive model. The study utilizes historical data from a cohort of computer science students, encompassing their GPAs and grades obtained in programming courses. The data is preprocessed, only including students' data that should have completed their study. The neural networks algorithm is then trained on this processed data, utilizing both the GPA and programming course grades as input features. The results obtained from our model indicate that the achieved accuracy was 0.75, which can be considered relatively low in our study's context. Our study underscores the need for further research and development to enhance the model's performance and address the limitations identified.**

Keywords*— *Length of study prediction, computer science students, GPA, programming course grades, neural networks algorithm.

I. INTRODUCTION

Students' academic performance is determined through the grades obtained from exams or assignments in the courses taken [18]. At the end of each semester, students will receive the grades for the courses taken in that semester. Students usually do well in some certain courses but there are also courses where they perform poorly or even fail due to various reasons. These grades will affect students' GPA which will determine whether they can graduate on time or not. In a study conducted by J. T. Denning, the higher GPA a student has, the more likely they are to meet certain academic requirements and thus graduate on-time[1]. Furthermore, students must repeat failing courses and get at least a minimum passing grade on those courses in order to graduate. Students will not be able to graduate if their GPA is below the minimum requirement even though they excel in some courses, or there are still courses they have to repeat even though their GPA is above the minimum requirement for graduation. In a study conducted by James Li about identifying the cause of risk of failure for medical students, GPA were found to be the best positive predictors of whether a student was likely to struggle and eventually fail[14].

Analyzing and predicting students' academic performance has shown successful results and benefits. It can be a potential revolutionary approach to complement the universities' methods to help students. The goal is to make out hidden patterns and prediction trends to find parts where students are lacking. This method has been used to solve other educational areas, such as student performance, dropout prediction and course selection.

There is some previous research that predicts student graduation based on certain courses. Francis Casillano's research was conducted using the decision tree algorithm [2]. A decision tree is an algorithm that divides all the data into nodes based on the purity of the classes [15]. His research is performed in six consecutive stages: data collection, data initial preparation, statistical analysis, data preprocessing, data mining implementation, and result evaluation. The decision tree algorithm will be applied using the software called Orange Data Mining, in which the decision tree will show the possibilities of whether a student will graduate on time or not based on their performance in the course. The algorithm also predicted that 52 students graduated on time, in which 49 were classified correctly while the other 3 were misclassified. The algorithm predicted that 20 students did not graduate on time, in which 49 were classified correctly while the other 3 were misclassified. In conclusion, the decision tree algorithm predictive model evaluation resulted in 88.9% accuracy.

In another study, Rohman A implemented the K-Nearest Neighbor algorithm to predict student degree accuracy [3]. The K-Nearest Neighbor algorithm is a classification method that groups new data based on the distance from the new data to some nearest data/neighbors. Tests tested with the K-Nearest-Neighbor algorithm yielded an accuracy score of 82.25% and an AUC score of 0.500. The cluster data k=2 yielded an accuracy score of 79.45% and an AUC score of 0.826. The cluster data k=3 yielded an accuracy score of 83.95% and an AUC score of 0.853. The cluster data k=4 yielded an accuracy score of 82.62% and an AUC score of 0.874. And the cluster data k=5 yielded an accuracy score of 85.15% and an AUC score of 0.888. It was concluded that the accuracy test of student completion using the K-Nearest Neighbor algorithm was affected by the number of clustering data, with clustering data k = 5 being the highest.

The studies mentioned above predict students graduate on time or not based on their GPA or grades in the courses taken. The contribution given in this study is to predict students' length of study based on their performances in certain courses. This way, people will be able to determine whether a student will graduate on time or not as well as the precise length of time it takes for a student to graduate.

Neural networks are particularly well-suited for predicting students' on-time graduation, as they can analyze enormous amounts of data and identify complex patterns and relationships that may be difficult to detect with any other methods [16-17]. In a study conducted by Mohammad Suhaimi N, and Abdul-Rahman S about predicting student's graduation time using different machine learning algorithms, Neural Network has the highest prediction accuracy above any other algorithms which is 95% [4]. This study uses Neural Networks prediction method to predict students' length of study with high level of accuracy. The result achieved in this study is to determine the model's level of accuracy and find strategies to further improve accuracy. Therefore, developers will know how to enhance their prediction in Neural Networks model.

This research paper will contain 5 sections. The first section will explain about problems of our research paper, which is the prediction of students' graduation rate based on their scores. The second section will explain some works that are related to our research. The third section will explain our method for solving the problem, by implementing the Neural Network model. The fourth section will contain the results and evaluations from our research. And finally, the fifth section will briefly summarize all our research.

II. Related Works

Neural Network (NN) is inspired by biological human brain that consists of up to 60 trillion interconnected set of neurons to perform network pattern of decision making [5]. The foundation of ANN is made up of a single layer of input, process (hidden layer) and output. It uses mathematical formulation sets to produce the most optimum result for a given problem. Neurons connect the fundamental building blocks of a neural network's information processing system. Every neuron contains more than one weight in addition to the adjusted weight, and each connection had its own weight. Links will travel from the input neuron to the following nodes. Each link in the chain is connected through the output by a number and a weight. All linked connections in this phase are called perceptron. Then, the input links will first be separately weighted, and their weights will be added up to create an activation function. This is called the Feed-forward (FF) neural network. After the FF approach comes the Back Propagation (BP) algorithm which is used to correct errors in the process. It is done by backtracking the process from the output layer to the hidden layer [6]. The goal of BP is to update the value of the weighted neurons and thus minimize errors for each output neuron and the whole network. Both FF and BP approaches are used together in training a neural network to achieve high accuracy in prediction tasks. This iterative process of forward propagation followed by backward propagation is repeated multiple epochs until the network's performance is satisfactory [19-20].

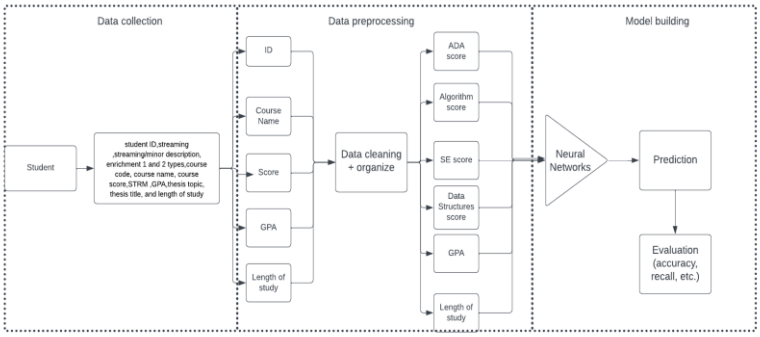
In a study conducted by Adekitan A and Salau O, they made a prediction model on students' graduation result based on their first three years of the students' GPA and their final GPA. The predictive model was developed in Konstanz Information Miner (KNIME) based data mining model and six main data mining algorithms (PNN, Random Forest, Decision Tree, Naïve Bayes, Tree Ensemble, and Logistic Regression) were applied in the KNIME model for predicting the class of grade of the final CGPA of the students at graduation. The result shows that the logistic regression algorithm has the highest accuracy (89.15%) while PNN algorithm has the lowest accuracy (85.895%) [7]. A study conducted by Widaningsih also compares several data mining algorithms (KNN, Naïve Bayes, SVM, C4.5) to predict students' final GPA and graduation time based on their gender and GPA from third to sixth semester. This study also implements cross validation (K-fold cross validation) to reduce computation time while maintaining estimation accuracy. The result shows that the Naïve Bayes algorithm has the best performance out of the four algorithms mentioned in this study. It achieved 76.79% accuracy with 23.17% error and AUC score of 0.850 [8].

Another similar study conducted by M T Sembiring and R H Tambunan predicts students' length of study (Late / On-time) based on their final GPA and some background features of the students. The prediction model uses Naïve Bayes algorithm and achieved only 70.83% accuracy. From this study, students with high GPA don't always graduate on time and students with average or low GPA might graduate on time [9]. Another study conducted by R. Ridman et al. that also predicts students' graduation rate based on their GPA by implementing Neural Networks algorithm achieved 98.27% accuracy. Such high accuracy can be achieved by implementing feed forward data input and backpropagation to improve the accuracy of the algorithm [10]. But, on the other hand, there is a study that predicts students' graduation based on their GPA and several background features that also uses neural networks algorithm (to be exact, ANN) that only reached around 0.62 (62%) accuracy, and even by using SMOTE enhanced ANN, it only reached at most 70.8% accuracy.[13]

While some studies use only a normal predictive model, a study conducted by Tampakas et al. uses a two-level classification to predict students' graduation time or if they didn't graduate. The A-level classifier is based on the student's background features, while the B-level classifier is based on the type and the final grade of the course taken. Result shows that the highest accuracy can be achieved by using 10NN in the A-level classifier and C4.5 algorithm in the B-level classifier which results in 78.73% accuracy. This indicates that the 10NN algorithm has the best performance in the A-level classifier while the C4.5 algorithm has the best performance in the B-level classifier [11]. Another study conducted by Ahmad et al. also modified their KNN model by implementing fuzzy technique to predict 1685 university students' graduation rate based on their GPA. This model also implements K-fold Cross Validation with k = 10 in data training. The result shows that this model achieved 77.35% accuracy from a total of 1138 of true positive data and 163 of true negative data [12].

III. METHODOLOGY

The aim of this research is to predict the length of study of students using their programming course grades, with the help of a neural network model. To achieve this goal, the following methodology was followed:



Data Collection

The data for this study was collected from the academic records of computer science university students which entered the university in 2018, 2019, and 2020. The dataset has a total of 120002 rows of data which includes the following features such as student ID, streaming (streaming type), streaming/minor description, enrichment 1 and 2 types, course code, course name, course score, STRM (semester), GPA, thesis topic, thesis title, and length of study up to even semester 2022/2023 (in semesters).

Data Preprocessing

The collected data was preprocessed to remove any missing or inconsistent data points. The first step is to remove students who have not completed their studies. The model should only use students' data that should have completed their study (joined 2018, graduated 2022). There are certain courses that have no grades which results in null value in the score column. Hence, several rows need to be removed to eliminate null values. There are multiple courses that a computer science student must take. But throughout the semester, not all students take the same course. Students with different streaming/minor will study different courses. So, only "Algorithm and Programming", "Data Structures", "Software Engineering", and "Algorithm Design and Analysis" course are selected as they are the required courses that computer science students must take regardless of the streaming/minor is. Then, there are also few features that don't have much effect on students' length of study such as their thesis topic and title, their enrichment details, streaming/minor, course code, and STRM. Therefore, these columns were deleted. And lastly, to make sure that all data values are number (integer/float), the dataset needs to be modified so that the features become student ID, Algorithm and Programming score, Data Structures score SE score, ADA score, GPA, and length of study. There are also some students that retake the same course which will result in duplicate score for a single course. So, only the highest score of the course is taken. Eventually, there are only 1078 rows of data that can be used to predict the length of study.

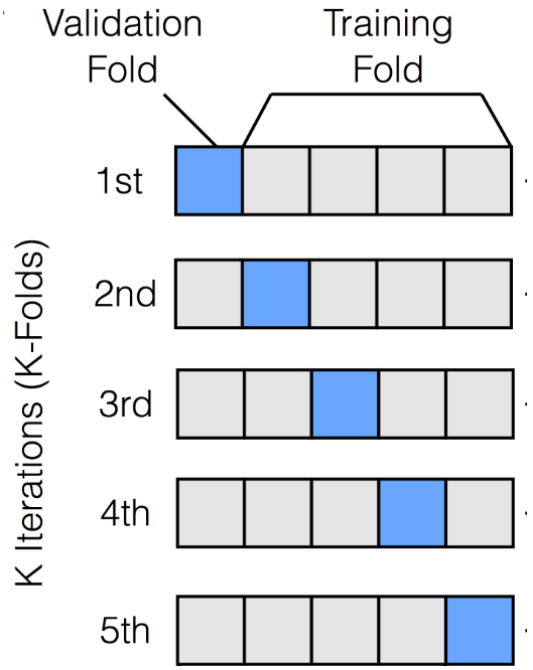
After the data is processed and cleaned, there is some data that must be removed. Students must study for at least 7 semesters (3.5 years), and they must pass every required course. This means that students with less than 7 semesters of studies are classified as "not graduated" because, students in this data are students that should have completed their study (graduated 2022) and having less than 7 semesters of study means that they have been dropped out / not continuing their study. Another case is that in the dataset, there are some students with 7 or 8 semesters of study. But some of their grades are still below the minimum passing grade, which means that they either have not graduated or dropped out or are not continuing their study. These students are also classified as "not graduated".

The dataset was then split into two parts: a training set and a testing set. The training set was used to train the neural network model, while the testing set was used to evaluate the performance of the model. The training set uses 80% of the dataset while the testing set uses 20% of the dataset.

Data Mining

This study uses Neural Networks algorithm to predict students' length of study based on their Algorithm and Programming and Data Structures course as well as their GPA. The Neural Networks model will have 5 input layers, 1-2 hidden layers, 1-10 / 90-100 hidden nodes for each layer, 0.01 learning rate, with "logistic" activation function, and max iteration of 1500. The goal of having several iterations with different number of hidden layers is to find out how many numbers of hidden layers have the best overall accuracy.

Another method that was utilized is implementing K-Fold Cross Validation. It is an evaluation method to improve the model's prediction on several unpredictable data. The illustration for K-Fold Cross Validation can be seen below:



The training dataset is partitioned into K - partitions. On the illustration above the K value is 5, which means that the training dataset is split equally into 5 partitions. On the first iteration, the first partition will be used as the validation dataset during the first iteration, while the remaining datasets will be used as the training dataset. The second iteration will use the second partition as the validation fold while the remaining datasets will be used as the training dataset. The third iteration will use the third partition as the validation fold, and so on. This process will repeat K times in total and the model will be trained throughout the whole process.

Interpretation and Evaluation

After predicting the result, the model will evaluate the model accuracy by using the accuracy score for each number of hidden layers. The most accurate number of hidden layers and nodes as well as the best accuracy score will be printed by the model.

IV. Result and Discussion

A. Experimental Preparation

Students' data which are classified as "not graduated" are deleted from the dataset because it will ruin the model's predicting accuracy. Moreover, the number of rows in the data where the "length of study" is less than 8 are very small which can lead to some of the data mentioned only goes into the training / testing dataset after splitting the data. The input will also be standardized first before splitting the data to improve accuracy.

B. Results Evaluations

The result of the neural networks (1-10 hidden neurons) model used to predict students' length of study is shown below:

2nd layer \ 1st layer	1	2	3	4	5	6	7	8	9	10
1	0.73	0.74	0.73	0.73	0.7	0.74	0.73	0.73	0.71	0.74
2	0.74	0.73	0.73	0.73	0.74	0.73	0.74	0.72	0.7	0.72
3	0.75	0.73	0.74	0.73	0.74	0.71	0.73	0.74	0.75	0.73
4	0.73	0.71	0.75	0.74	0.74	0.72	0.75	0.74	0.74	0.74
5	0.72	0.73	0.73	0.74	0.74	0.73	0.73	0.73	0.75	0.74
6	0.74	0.72	0.72	0.72	0.74	0.73	0.74	0.73	0.73	0.74
7	0.73	0.73	0.74	0.73	0.74	0.73	0.73	0.75	0.75	0.73
8	0.74	0.73	0.71	0.73	0.73	0.74	0.73	0.72	0.73	0.75
9	0.74	0.72	0.72	0.71	0.73	0.73	0.74	0.75	0.72	0.73
10	0.74	0.73	0.7	0.74	0.74	0.73	0.74	0.72	0.74	0.73

And the result of 90 – 100 hidden neurons are shown below:

2nd layer \ 1st layer	90	91	92	93	94	95	96	97	98	99	100
90	0.73	0.73	0.74	0.73	0.73	0.74	0.73	0.73	0.73	0.58	0.73
91	0.75	0.75	0.6	0.74	0.74	0.74	0.73	0.75	0.74	0.73	0.74
92	0.74	0.73	0.75	0.73	0.65	0.72	0.6	0.73	0.75	0.74	0.74
93	0.72	0.74	0.74	0.72	0.74	0.73	0.74	0.73	0.72	0.74	0.74
94	0.73	0.61	0.73	0.73	0.74	0.73	0.74	0.63	0.75	0.73	0.73
95	0.74	0.73	0.72	0.75	0.58	0.6	0.7	0.73	0.73	0.73	0.73
96	0.72	0.74	0.73	0.72	0.71	0.73	0.72	0.73	0.73	0.74	0.74
97	0.74	0.73	0.73	0.74	0.73	0.73	0.7	0.75	0.74	0.73	0.74
98	0.61	0.74	0.73	0.74	0.74	0.73	0.74	0.74	0.74	0.73	0.73
99	0.74	0.74	0.73	0.75	0.74	0.74	0.74	0.74	0.6	0.74	0.72
100	0.73	0.72	0.73	0.56	0.73	0.73	0.71	0.74	0.73	0.72	0.75

From the tables above, it seems that using more hidden nodes does not improve the model's prediction accuracy. This indicates that another approach is needed to improve the accuracy of the model.

With mostly the same model configuration (100 hidden neurons for each hidden layer, 1500 iteration, 0.01 learning rate, activation = 'logistic'), K-Fold Cross Validation is implemented to the model. The table below shows the result of implementing the K-Fold Cross Validation in the students' dataset with K-value = 2 – K-value = 20:

K-value	Accuracy
2	0.72
3	0.62
4	0.74
5	0.58
6	0.74
7	0.75
8	0.74
9	0.73
10	0.74
11	0.73
12	0.57
13	0.6
14	0.7
15	0.68
16	0.74
17	0.75
18	0.72
19	0.75
20	0.73

The table above shows that the K-Fold Cross Validation still has not improved the accuracy of the model and has the best accuracy of 0.75 on K = 7, K = 17, and K = 19.

C. Results Analysis

From the results shown above, it can be observed that the initial model accuracy scores range from 0.7 to 0.75 through different configurations. It is interesting to note that increasing the number of neurons in the hidden layer does not consistently improve the model's accuracy. For example, when the model has 10 neurons in the hidden layer, the accuracy drops to 0.7, which is lower compared to the configuration with 8 neurons. This indicates there might be a point of diminishing returns when it comes to the model architecture's complexity.

Even though the data has gone various processing such as data standardization, the highest accuracy that this neural network model can reach is only around 0.75 (75%), which is not really high. This means that the model does not predict students' length of study very well by using their GPA and some core programming courses scores. Another reason that the accuracy is not very high is that the data is unpredictable, where some students graduate on time with grades just right above the passing grade while some

students have high grades in certain scores but still do not graduate on time.

To adapt the model from the unpredictable data, K-Fold Cross Validation was implemented, but still did not manage to increase the model's accuracy. This means that the training dataset does not describe the test dataset very well, which makes the model that is trained based on the training dataset unable to predict the testing dataset precisely.

V. Conclusion and Future Works

In conclusion, our study demonstrates the potential of the neural network algorithm in predicting the length of study for Computer Science students. However, further optimization and evaluation are required to enhance the accuracy and robustness of the model, and to assess its applicability in different contexts. Additionally, it is crucial to assess its applicability in diverse contexts to ensure its generalizability.

Our findings underscore the need for continued research and development to refine the model's performance. It is important to acknowledge that these results are based on a specific dataset, and the effectiveness of the neural network model may vary when applied to different datasets or influenced by additional factors.

Overall, our study lays the foundation for future investigations in this domain and highlights the potential of neural network algorithms in predicting study durations for Computer Science students. Continued efforts in research and refinement will contribute to the development of accurate and reliable models that can effectively assist educational institutions in resource planning and student support.

References

[1] J. T. Denning, E. R. Eide, K. J. Mumford, R. W. Patterson, and M. Warnick, "Why Have College Completion Rates Increased?," Am Econ J Appl Econ, vol. 14, no. 3, pp. 1–29, Jul. 2022, doi: 10.1257/app.20200525.

[2] J. Co and N. Francis Casillano, "Predicting On-time Graduation based on Student Performance in Core Introductory Computing Courses using Decision Tree Algorithm," Jurnal Pendidikan Progresif, vol. 11, no. 3, pp. 650–658, 2021, doi: 10.23960/jpp.v11.i3.202116.

[3] A. Rohman, "MODEL ALGORITMA K-NEAREST NEIGHBOR (K-NN) UNTUK PREDIKSI KELULUSAN MAHASISWA," Neo Teknika, vol. 1, no. 1, Mar. 2015, doi: 10.37760/neoteknika.v1i1.350.

[4] N. Mohammad Suhaimi, S. Abdul-Rahman, S. Mutalib, N. H. Abdul Hamid, and A. Hamid, "Review on Predicting Students' Graduation Time Using Machine Learning Algorithms," International Journal of Modern Education and Computer Science, vol. 11, no. 7, pp. 1–13, Jul. 2019, doi: 10.5815/ijmecs.2019.07.01.

[5] A. G. Farizawani, M. Puteh, Y. Marina, and A. Rivaie, "A review of artificial neural network learning rule based on multiple variant of conjugate gradient approaches," in *Journal of Physics: Conference Series*, 2020. doi: 10.1088/1742-6596/1529/2/022040.

[6] A. G. Farizawani, M. Puteh, Y. Marina, and A. Rivaie, "A review of artificial neural network learning rule based on multiple variant of conjugate gradient approaches," in *Journal of Physics: Conference Series*, 2020. doi: 10.1088/1742-6596/1529/2/022040.

[7] 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.

[8] S. Widaningsih, "PERBANDINGAN METODE DATA MINING UNTUK PREDIKSI NILAI DAN WAKTU KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA DENGAN ALGORITMA C4,5, NAÏVE BAYES, KNN DAN SVM," *Jurnal Tekno Insentif*, vol. 13, no. 1, pp. 16–25, Apr. 2019, doi: 10.36787/jti.v13i1.78.

[9] M. T. Sembiring and R. H. Tambunan, "Analysis of graduation prediction on time based on student academic performance using the Naïve Bayes Algorithm with data mining implementation (Case study: Department of Industrial Engineering USU)," *IOP Conf Ser Mater Sci Eng*, vol. 1122, no. 1, p. 012069, Mar. 2021, doi: 10.1088/1757-899X/1122/1/012069

[10] R. Ridwan, H. Lubis, and P. Kustanto, "Implementasi Algoritma Neural Network dalam Memprediksi Tingkat Kelulusan Mahasiswa," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, p. 286, Apr. 2020, doi: 10.30865/mib.v4i2.2035.

[11] V. Tampakas, I. E. Livieris, E. Pintelas, N. Karacapilidis, and P. Pintelas, "Prediction of Students' Graduation Time using a Two-Level Classification Algorithm," 2019, pp. 553–565. doi: 10.1007/978-3-030-20954-4_42.

[12] I. Ahmad, H. Sulistiani, and H. Saputra, "The Application Of Fuzzy K-Nearest Neighbour Methods for A Student Graduation Rate," *Indonesian Journal of Artificial Intelligence and Data Mining*, vol. 1, no. 1, p. 47, Nov. 2018, doi: 10.24014/ijaidm.v1i1.5654.

[13] A. Yaqin, M. Rahardi, and F. F. Abdulloh, "Accuracy Enhancement of Prediction Method using SMOTE for Early Prediction Student's Graduation in XYZ University," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 6, 2022, doi: 10.14569/IJACSA.2022.0130652.

[14] J. Li, R. Thompson, and B. Shulruf, "Struggling with strugglers: using data from selection tools for early identification of medical students at risk of failure," BMC Med Educ, vol. 19, no. 1, p. 415, Dec. 2019, doi: 10.1186/s12909-019-1860-z.

[15] K. Kim, "A hybrid classification algorithm by subspace partitioning through semi-supervised decision tree," *Pattern Recognition*, vol. 60, 2016, doi: 10.1016/j.patcog.2016.04.016.

[16] X. Li, Y. Zhang, H. Cheng, M. Li, and B. Yin, "Student achievement prediction using deep neural network from multi-source campus data," *Complex and Intelligent Systems*, vol. 8, no. 6, 2022, doi: 10.1007/s40747-022-00731-8.

[17] O. A. Montesinos López, A. Montesinos López, and J. Crossa, "Fundamentals of Artificial Neural Networks and Deep Learning," in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, 2022. doi: 10.1007/978-3-030-89010-0_10.

[18] S. K. F. Briones, R. J. R. Dagamac, J. D. David, and C. A. B. Landerio, "Factors Affecting the Students' Scholastic Performance: A Survey Study," *Indonesian Journal of Educational Research and Technology*, vol. 2, no. 2, 2022, doi: 10.17509/ijert.v2i2.41394.

[19] A. R. Yessa and M. Hardjianto, "Prediction of Water Use Using Backpropagation Neural Network Method and Particle Swarm Optimization," *bit-Tech*, vol. 2, no. 3, 2020, doi: 10.32877/bt.v2i3.158.

[20] A. Wanto *et al.*, "Optimization of Performance Traditional Back-propagation with Cyclical Rule for Forecasting Model," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 22, no. 1, 2022.