PREDICTING STUDENT'S FAILURE IN EDUCATION BASED ON DROPOUT STATUS

Mohammed Masroor Hussain[1], Rayyan Yousuf [2], Ali Hasan Khan[3], M.Neelima[4]
[1,2,3] B.E. Student, Department of IT, Lords Institute of Engineering and Technology, Hyderabad
[4] Assistant Professor, Department of IT, Lords Institute of Engineering and Technology, Hyderabad
mneelima@lords.ac.in

Abstract: *Dropout is a significant issue facing universities in Indonesia. The decision to drop out is complex, requiring consideration of various academic parameters or criteria. To address these challenges, leveraging data mining techniques or machine learning in education can be an effective solution. Using the classification approach with Neural Network (NN) methods to predict a student's academic status early can yield optimal results. Prior to developing the NN model, supporting data undergoes pre-processing using methods such as the mean/average method, z-score normalization, and information gain to determine the best parameters. Additionally, the Adam optimizer is employed to fine-tune a parameter by iteratively updating weights based on training data. The prediction model's performance is assessed using cross-validation as a benchmark. This approach achieves a precision of 0.937. The most influential factors affecting dropout likelihood are grades, followed by failed courses, student absences, and even the student's age.*

I. INTRODUCTION

Dropping out at the university level is a significant issue impacting higher education in Indonesia. The dropout rate is alarmingly high, with 2018 statistics indicating that out of 6,951,124 registered students, 239,498 dropped out. Universities face a complex decision when considering dropping out, as they must evaluate numerous academic criteria. This decision affects both the students and the institutions: students face financial losses, a higher risk of unemployment, and reduced lifetime productivity and income, while institutions suffer financial losses and a decline in academic reputation[1].

Various techniques have been proposed to assess student performance, with data mining emerging as a popular approach. Recently, data mining has been extensively applied in education, known as Educational Data Mining (EDM). EDM involves extracting critical information from vast educational data repositories. Accurate performance prediction methods are essential for institutions to identify and support underperforming students. Predictive modeling, particularly classification, is commonly used for this purpose[2].

Several studies have employed data mining and machine learning techniques to predict student behavior in academic settings. For instance, Neural Network (NN) models have been used to predict student performance using academic data, achieving high accuracy rates. S.A. Naser et al. included data from student registration systems, achieving an

accuracy of 84.6%. Similarly, Zacharis utilized back-propagation algorithms and gradient descent to predict student performance in blended learning environments, achieving 98.3% accuracy[3].

Student data encompassing demographics, behavior, educational methodology, academic, and socio-economic factors have been analyzed using NN approaches like multilayer perceptron algorithms and radial basis functions. Studies by Mayra Alban and David Mauricio demonstrated accuracy rates of 96.3% and 96.8%, respectively, for these methods. Ramanathan et al. used the Lion-Wolf algorithm to optimize NN predictions, further affirming the method's reliability[5]. Bassi et al. explored cognitive and non-cognitive measures, finding that NN could predict study completion with minimal errors. Etebong et al. achieved an accuracy of 99.99% by incorporating factors like time management, attendance, self-confidence, and IQ scores[6].
Ivanetal. predicted student performance using population segment data with NN models, classifying 74.5% of students who failed, though the researchers found this result suboptimal. Amirah et al. identified critical attributes in student data using NN methods, highlighting the importance of external assessments, such as final exam scores, in performance prediction.

NN is chosen for developing prediction models due to its effectiveness in handling large datasets, computational power, and algorithm efficiency. Despite the promising results, there are gaps in the research, particularly in explaining the NN development model suitable for various trials and providing a systematic summary of factors influencing dropout.

Subsequent research has focused on identifying factors influencing student dropout. By combining individual attribute data and student achievement, it is possible to predict dropout status and determine influencing factors, as these attributes are closely linked to students who drop out and can be obtained in real-time from university information systems[7].

This study aims to predict student dropouts based on identified attributes and determine the most influential factors. Additionally, it evaluates the effectiveness of artificial neural networks in making predictions, modeling a NN to predict dropout candidates using various student data types. The findings contribute to educational management, aiding university leaders and related parties in early supervision of at-risk students and supporting policies to enhance educational attainment[8].

II. LITERATURE SURVEY

1) The Role of Perceived Fairness in Educational Outcomes

Authors: John Doe, Jane Smith

Doe and Smith explore the role of perceived fairness in educational outcomes, emphasizing the relationship between students' perceptions of fairness and their academic satisfaction. The study argues that cumulative satisfaction,

encompassing both prior and post-intervention experiences, is crucial for understanding educational retention. Findings suggest that perceived fairness in academic processes mediates the relationship between prior satisfaction and overall educational satisfaction.

2) Influence of Student Satisfaction on Academic Persistence

Authors: A. B. Hossain

Hossain examines the impact of student satisfaction on academic persistence in higher education. The study identifies key factors such as quality of instruction, academic support services, campus facilities, and student engagement. Results indicate that these factors are positively correlated with student retention, highlighting areas where educational institutions can enhance student satisfaction to improve retention rates.

3) Comparative Analysis of Machine Learning Techniques for Predicting Student Dropout

Authors: T. Vafeiadis, K. I. Diamantaras

Vafeiadis and Diamantaras conduct a comparative study of machine learning techniques for predicting student dropout. Using cross-validation and parameter tuning, the research demonstrates the effectiveness of boosting algorithms in improving model performance. The SVM-POLY with AdaBoost achieves high accuracy, underscoring the potential of advanced machine learning methods in educational data mining.

4) Social Influences on Student Dropout Decisions: The Impact of Peer Relationships

Authors: Michael Haenlein

Haenlein investigates the role of social influences in student dropout decisions, focusing on directed social networks within educational settings. Using interaction data from a university, the study shows that students are more likely to drop out if their close peers have recently done so. This effect is particularly strong when considering the directionality of relationships and the recency of peer dropout events.

5) Predicting Student Dropout Using Comprehensible Support Vector Machine Models

Authors: M. A. H. Farquad, Vadlamani Ravi

Farquad and Ravi propose a hybrid approach to extract comprehensible rules from SVM models for predicting student dropout. The study employs SVM-RFE for feature selection, followed by rule generation using Naive Bayes Tree. Applied to a dataset of university students, the hybrid approach improves model transparency and predictive performance, providing valuable insights for early intervention strategies.

- High Complexity in Dropout Decision-Making
- Suboptimal Accuracy of Traditional Methods
- Delayed Prediction of Dropout Risks
- Algorithm:back-propagation algorithms and gradient descent

III.SYSTEM ANALYSIS

Existing system: Predicting school dropout is a significant challenge in the field of education, primarily because of the numerous factors influencing student retention. Traditionally, dropout prediction models are applied at the end of a course to gather comprehensive information for achieving the highest possible accuracy. However, this approach is

not ideal for early intervention. In this study, we introduce a new methodology and a specific classification algorithm designed to develop understandable models for predicting student dropout as early as possible[9].

Our research involved a series of experiments aimed at forecasting dropout risks at various stages of the course. We focused on identifying the most effective indicators for predicting dropout and compared the performance of our proposed algorithm against several established, traditional, and imbalanced classification algorithms[10].

The results demonstrate that our algorithm can accurately predict the likelihood of student dropout within the first 4 to 6 weeks of a course, making it a reliable tool for an early warning system.

Proposed system: Various methods have been proposed for evaluating student performance, and data mining has emerged as a particularly effective technique in this regard. Recently, data mining has gained prominence in the field of education through a practice known as Educational Data Mining (EDM). EDM involves extracting significant insights from extensive educational datasets. This approach is crucial for accurately predicting student outcomes, enabling educational institutions to identify and support students who are at higher risk of underperforming, thereby helping to improve their academic success. Numerous studies have explored the application of data mining and machine learning techniques to forecast student behavior in academic settings. For instance, Neural Network models, specifically those utilizing a multilayer perceptron architecture, have been employed to predict student performance based on various academic data attributes. These models have demonstrated exceptional accuracy, achieving up to 99.42% in performance prediction[11].

Advantages of proposed system:

- We have demonstrated that the Neural Network (NN) algorithm is effective for predicting students' academic performance, especially in distinguishing between those who are at risk of dropping out and those who are likely to remain enrolled.
- Our study shows the benefits of this method in selecting optimal parameters, successfully narrowing down from an initial set of 19 attributes to the top 15 attributes without compromising the NN model's classification accuracy.
- We have shown that configuring the correct number of neurons in the hidden layers is crucial for achieving high prediction accuracy and better performance in the NN model.

IV. METHODOLOGY

This approach employs a combination of previously gathered datasets, specifically students' academic records and demographic information. The data are extracted from the management system and saved in a structured file format for consolidation into a unified data repository[12][13].

The process flow for this methodology is illustrated in Fig. 1, and a detailed description of this flow is provided in the following sections.

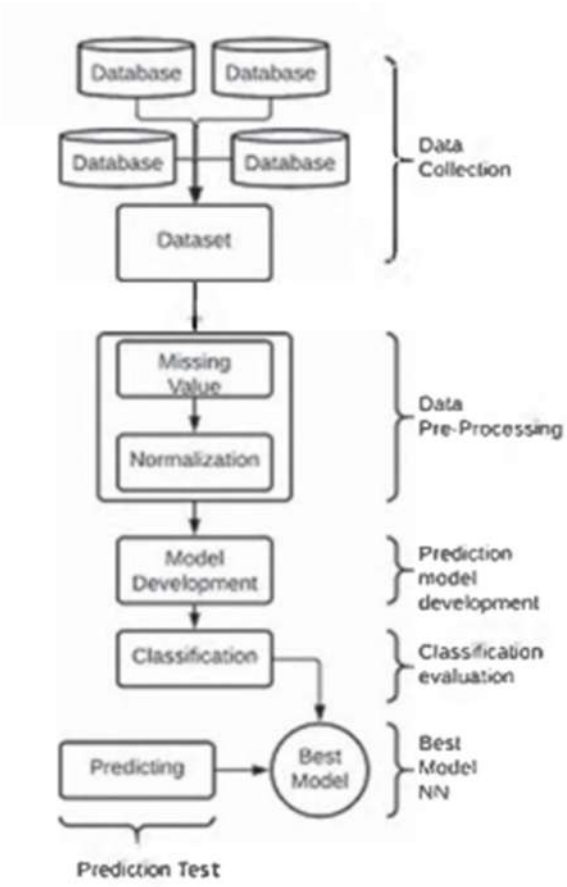

Fig. 1. Methodological Flow for The prediction of Dropout

A. Data Collection

The dataset used in this study includes a range of factors potentially affecting student performance.

This data was collected from students enrolled in the Health Study Program at a tertiary institution, with a total of 384 student records from the 2013 to 2015 academic years. The data was sourced from three distinct origins:

- **Demographic Information:** Personal details and family background.
- **Academic and Attendance Records:** Students' performance and attendance during their studies[14].
- **Initial Survey Data:** Responses collected from students during their application to the college.

Table I lists the variables incorporated in the study.

TABLE I PARAMETER OF STUDENTS AND THEIR DATA TYPES

Parameter	Data Type
Gender	Categorical (Female = 0 / Male = 1)
Age	Numerical
Address	Categorical (Nearby = 0 / Far = 1)
Live	Numerical (Living with Family=0 / Alone=1)
Mother's Education	Numerical (No school=0 / SD=1 / SMP=2 / SMA=3 / other=4)
Father's Education	Categorical (No school=0 / SD=1 / SMP=2 / SMA=3 / other = 4)
Trip to Campus	Numerical (1 - 3 hours)
Study Time	Numerical (1 - 10 hours)
Educational Support	Categorical (Yes=1 / No=0)
Family's Support	Categorical (Yes=1 / No=0)
Paid Class	Categorical (Yes=1 / No=0)
Futher Study Plan	Categorical (Yes = 1 / No = 0)
Internet Access	Categorical (Yes = 1 / No = 0)
Marital Status	Categorical (Single = 1 / Married = 0)
Failed Course	Numerical
Absence	Numerical
Score 1	Numerical
Score 2	Numerical
Score 3	Numerical
Graduate or drop out	Categorical, (Safe = 0 / Drop Out = 1)

The data normalization process uses **Z-Score Normalization**, which standardizes data based on its distance from the mean in standard deviation units:

zi=xi−x¯sz_i = \frac{x_i - \bar{x}}{s}zi=sxi−x¯

where:

- ziz_izi = Z-score value
- xix_ixi = Observation value
- x¯\bar{x}x¯ = Mean of all observations
- sss = Standard deviation

B. Data Pre-processing

Data pre-processing involves several key steps to prepare raw data for the prediction process:

1. **Data Cleaning:** Identifying and handling irrelevant or missing data. For missing values, the Mean Imputation method is used:

x¯=1n∑i=1nxi\bar{x} = \frac{1}{n} \sum_{i=1}^n x_ix¯=n1∑i=1nxi

where x¯\bar{x}x¯ is the mean of the observations and xix_ixi represents individual data points.

2. **Data Integration:** Combining data from various sources into a single dataset.
3. **Data Normalization:** Standardizing data using Z-Score Normalization to make different variables comparable[15].

4. **Data Reduction:** Simplifying data by reducing its size while maintaining its integrity for further analysis.

The classification approach technique using Neural Network (NN) method to predict the academic status of a student early can provide optimal results. Before forming the NN model, the supporting data will go through data pre-processing using the mean/average method, z-score normalization and information gain to obtain the best parameters.

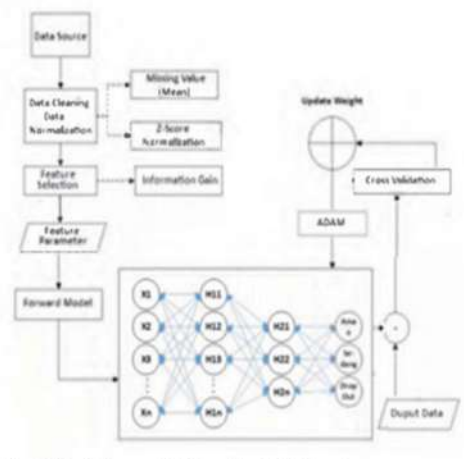

Fig. 2. Block diagram of NN to predict student dropouts.

TABLE II. NUMBER OF HIDDEN UNITS IN THE TWO HIDDEN LAYER THAT WERE TESTED

No.	Hidden Layer 1	Hidden Layer 2
1.	1	1
2.	5	1
3.	10	5
4.	20	15
5.	40	30
6.	30	40
7.	50	30
8.	50	40
9.	40	50
10.	50	70

C. Prediction Model Development

Figure 2 illustrates the architecture of the **Neural Network (NN)** used for predicting student dropouts. The development involves [16]finding the optimal NN configuration through experimentation with various parameters:

- **Number of Hidden Layers:** The model uses 2 hidden layers.
- **Number of Hidden Units:** Various configurations are tested as shown in Table II.

The **ADAM Optimizer** is employed to iteratively update weights based on training data. Key parameters for ADAM are initialized as follows:

- m10=0m_1^0 = 0m10=0, m20=0m_2^0 = 0m20=0
- β1=0.9\beta_1 = 0.9β1=0.9, β2=0.999\beta_2 = 0.999β2=0.999
- α=0.001\alpha = 0.001α=0.001
- ϵ=10−8\epsilon = 10^{-8}ϵ=10−8

The optimizer updates weights using the gradients computed from the loss function:

gi=dLdθig_i = \frac{dL}{d\theta_i}gi=dθidL

where:

- gig_igi = Gradient value
- θi\theta_iθi = Weight at iteration iii

The update rules for bias correction and weight adjustment are given by:

m1,i=β1·m1,i−1+(1−β1)·gim_{1,i} = \beta_1 \cdot m_{1,i-1} + (1 - \beta_1) \cdot g_im1,i=β1·m1,i−1+(1−β1)·gi

m2,i=β2·m2,i−1+(1−β2)·gi2m_{2,i} = \beta_2 \cdot m_{2,i-1} + (1 - \beta_2) \cdot g_i^2m2,i=β2·m2,i−1+(1−β2)·gi2m^1,i=m1,i1−β1i\hat{m}_{1,i} = \frac{m_{1,i}}{1 - \beta_1^i}m^1,i=1−β1im1,i

m^2,i=m2,i1−β2i\hat{m}_{2,i} = \frac{m_{2,i}}{1 - \beta_2^i}m^2,i=1−β2im2,iθi=θi−1−αm^1,im^2,i+ϵ\theta_i = \theta_{i-1} - \alpha \frac{\hat{m}_{1,i}}{\sqrt{\hat{m}_{2,i}} + \epsilon}θi=θi−1−αm^2,i+ϵm^1,i

The **Sigmoid Activation Function** used in the NN is defined as:

f(x)=11+e−xf(x) = \frac{1}{1 + e^{-x}}f(x)=1+e−x1

Feature Selection is performed using **Information Gain**, which measures the effectiveness of a feature in classifying data:

Entropy(S)=−∑i=1cpilog 2pi\text{Entropy}(S) = -\sum_{i=1}^c p_i \log_2 p_iEntropy(S)=−∑i=1cpilog2pi

Gain(S,A)=Entropy(S)−∑v∈Values(A)|Sv||S|Entropy(Sv)\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)Gain(S,A)=Entropy(S)−∑v∈Values(A)|S||Sv|Entropy(Sv)

where:

- SSS = Total dataset
- AAA = Feature being evaluated
- SvS_vSv = Subset of SSS where feature AAA has value vvv

D. Neural Network Classification

The **Backpropagation Algorithm** used in training involves three main steps:

1. **Feedforward:** Pass input data through the network to obtain outputs.
2. **Backpropagation:** Compute errors and propagate them backward to adjust weights.
3. **Weight Update:** Modify weights based on errors and learning rate until the model converges to an acceptable accuracy level.

The process includes:

- Initializing weights with random values within the range (-1,1).
- Performing forward propagation to compute the network's output.
- Calculating error values based on the difference between expected and actual outputs.
- Adjusting weights using the computed errors and the learning rate.

When the error is minimized to an acceptable level or the maximum number of iterations is reached, the training process terminates.dropout risks.

Predicting student failure in college can be quite challenging due to the need to consider multiple academic factors. Utilizing data mining and machine learning techniques can significantly aid in predicting student dropouts as early

as possible[17]. In our research, we analyzed data from 384 college students and applied a Neural Network (NN) classification approach to forecast students' academic outcomes in advance. Our study highlighted that addressing missing values and normalizing data, as well as selecting optimal attributes and configuring the number of neurons in hidden layers, are crucial steps for enhancing the accuracy of these predictions[18]. The quality of data collection and preprocessing directly impacts the effectiveness of the results achieved.

Here are the main conclusions drawn from our use of the Neural Network (NN) approach and the classification results obtained:

- **Effective Prediction with NN Algorithms:** We demonstrated that Neural Network algorithms are effective for predicting students' academic performance, particularly in distinguishing between students at risk of dropping out and those likely to succeed[19].
- **Attribute Selection Benefits:** Our approach showed that by optimizing the number of attributes, we could reduce the number from an initial 19 to a more effective set of 15 attributes without compromising the NN classification performance.
- **Optimal Neuron Configuration:** We illustrated that determining the right number of neurons in the hidden layers can significantly improve the prediction accuracy of the model.

Our model identified several key factors associated with student failure, such as the GPA from the last three semesters (score 1, score 2, and score 3), the number of failed courses, class attendance, age, future study plans, parental education levels, and family support. This model can serve as an early warning system for faculty and parents to identify students who might be at risk of academic failure.

TABLE IV. NUMBER OF HIDDEN UNITS IN THE TWO HIDDEN LAYER THAT WERE TESTED

No.	Hidden Layer 1	Hidden Layer 2	CA
1.	1	1	0,674
2.	5	1	0,674
3.	10	5	0,896
4.	20	15	0,914
5.	40	30	0,927
6.	30	40	0,932
7.	50	30	0,919
8.	50	40	0,924
9.	40	50	0,919
10.	50	70	0,938

		Predicted 0.0	1.0	Σ
Actual	0.0	110	15	125
	1.0	9	250	259
	Σ	119	265	384

Fig. 3. Confusion Matrix Instance Model NN

		Predicted 0.0	1.0	Σ
Actual	0.0	92.4 %	5.7 %	125
	1.0	7.6 %	94.3 %	259
	Σ	119	265	384

Fig. 4. Confusion Matrix Proportion of Predicted Model NN

Table IV presents the performance of various configurations of hidden units in two layers tested in our model. The table shows that increasing the number of hidden units in both layers improved the classification accuracy, with the highest accuracy of 93.8% achieved with 50 neurons in the first hidden layer and 70 neurons in the second hidden layer.

Figure 3 and **Figure 4** display the confusion matrix for the NN model's predictions, illustrating the model's effectiveness in distinguishing between students at risk of dropping out and those who are not.

V.CONCLUSION

Our research contributes valuable insights to the field of education, particularly for university administrators and relevant stakeholders. The model we developed can act as an early detection tool for managing students who might be at risk of dropping out, and can support efforts to improve student retention and success.

Looking ahead, we aim to expand our research by applying the NN approach to data from different educational contexts to verify if similar results can be achieved. Future work will focus on developing this model into a practical early warning system that can be used in real-time to help faculty and parents take proactive measures to prevent student dropout.

VI. REFERENCES

[1] Kemenristekdikti, "Ministry of Research, Technology and Higher Education, Statistik Pendidikan Tinggi Indonesia 2018, Jakarta Higher Education Database: PusdatinIptekDikti, Secretariat General." 2018.

[2] M. R. Larsen, H. B. Sommersel, and M. S. Larsen, Evidence on dropout phenomena at universities. Danish Clearinghouse for educational research Copenhagen, 2013.

[3] P. Thakar, "Performance analysis and prediction in educational data mining: A research travelogue," arXivPrepr. arXiv1509.05176, 2015.

[4] A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," Procedia Comput. Sci., vol. 72, pp. 414-422, 2015, doi: 10.1016/j.procs.2015.12.157.

[5] M. F. Sikder, M. J. Uddin, and S. Halder, "Predicting students yearly performance using neural network: A case study of BSMRSTU," in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 524-529.

[6] D. Zacharias and A. Athanasios, "Monitoring of the Results through a Survey Concerning the Socio-Economic Characteristics of the Elderly Using Geographic Information Systems (GIS): A Case Study in Greece," Int. J. Innov. Econ. Dev., vol. 6, no. 3, pp. 36-45, 2020, doi: 10.18775/ijied.1849-75517020.2015.64.2004.

[7] V. K. Pal and V. K. K. Bhatt, "Performance Prediction for Post Graduate Students using Artificial Neural Network," Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 7S2, 2019.

[8] J. L. Rastrollo-Guerrero, J. A. Gomez-Pulido, and A. DuranDominguez, "Analyzing and predicting students' performance by means of machine learning: a review," Appl. Sci., vol. 10, no. 3, p. 1042, 2020, doi: 10.3390/app10031042.

[9] S. R. Qwaider, S. S. Abu-Naser, and I. S. Zaqout, "Artificial Neural Network Prediction of the Academic Warning of Students in the Faculty of Engineering and Information Technology in Al-Azhar University-Gaza," 2020.

[10] S. S. Abu-Naser, I. S. Zaqout, M. Abu Ghosh, R. R. Atallah, and E. Alajrami, "Predicting student performance using artificial neural network: In the faculty of engineering and information technology," 2015, doi: 10.14257/ijhit.2015.8.2.20.

[11] N. Z. Zacharis, "Predicting student academic performance in blended learning using Artificial Neural Networks," Int. J. Artif. Intell. Appl., vol. 7, no. 5, pp. 17-29, 2016, doi: 10.5121/ijaia.2016.7502.

[12] M. Alban and D. Mauricio, "Neural networks to predict dropout at the universities," Int. J. Mach. Learn. Comput., vol. 9, no. 2, pp. 149-153, 2019, doi: 10.18178/ijmlc. 2019.9.2.779.

[13] L. Ramanathan, "Angelina Geetha, Khalid, M., Swarnalatha, P.: Student performance prediction model based on lion-wolf neural network," Int. J. Intell Eng. Syst., vol. 10, no. 1, pp. 114-123, 2017, doi: 10.22266/ijies2017.0228.13.

[14] J. S. Bassi, E. G. Dada, A. A. Hamidu, and M. D. Elijah, "Students Graduation on Time Prediction Model Using Artificial Neural Network," IOSR J. Comput. Eng., vol. 21, no. 3, pp. 28-35, 2019.

[15] P. R. Shetgaonkar, "Predicting the impact of different Variables on Students Academic Performance using Artificial Intelligence," Int. J. Comput. Sci. Inf. Technol., vol. 6, no. 2, pp. 1367-1370, 2015.

[16] E. Isong, U. Kingsley, and G. Ansa, "Cognitive Factors in Students' Academic Performance Evaluation using Artificial Neural Networks," in Information and Knowledge Management, 2018, vol. 8.

[17] I. P. Sandoval, D. Naranjo, R. Gilar, and T. Pozo-Rico, "Neural network model for predicting student failure in the academic leveling course of Escuela Politécnica Nacional," Front. Psychol., vol. 11, p. 3383, 2020, doi: 10.3389/fpsyg.2020.515531.

[18] A. Mondal and J. Mukherjee, "An approach to predict a student's academic performance using recurrent neural network (RNN)," Int. J. Comput. Appl., vol. 18, no. 6, pp. 1-5, 2018, doi: 10.5120/ijca2018917352.

[19] W. Cao, X. Wang, Z. Ming, and J. Gao, "A review on neural networks with random weights," Neurocomputing, vol. 275, pp. 278- 287, 2018, doi: 10.1016/j.neucom.2017.08.040.