2024 IEEE Conference on Artificial Intelligence (CAI)

Prediction of Students' Academic Progression using Machine Learning

Lien Nguyen
Centre for Teaching, Research and Scholarship
Sydney Institute of Higher Education
Sydney, Australia
lien.nguyen@sydneyinstitute.edu.au

Anh Nguyen
School of Electrical Engineering
Hanoi University of Science and Technology
Hanoi, Vietnam
anh.nguyenvan1@hust.edu.vn

Jack Jia
Centre for Teaching, Research and Scholarship
Sydney Institute of Higher Education
Sydney, Australia
jack.jia@sydneyinstitute.edu.au

Steve Ling
School of Electrical and Data Engineering
University of Technology Sydney
Sydney, Australia
Steve.Ling@uts.edu.au

Nigel Finch
Centre for Teaching, Research and Scholarship
Sydney Institute of Higher Education
Sydney, Australia
nigel.finch@sydneyinstitute.edu.au

***Abstract*—This paper is concerned with the development of an innovative method for predicting students' academic performance and progression in their first semester of studies at higher education institutions in Australia. This is an essential factor contributing to the improvement of student retention which is considered as one of the most challenging problems within higher education sector, especially for private institutions. An artificial neural network was developed and trained by a training set to learn patterns of a group of 200 commencing students at Sydney Institute of Higher Education (SIHE). The neural network is subsequently tested with a testing set which consists of data from another group of 100 commencing students to examine the ability of predicting students' achievement of learning outcome in their first semester at SIHE. With the best classification results of 85% Sensitivity and 69% Specificity on the training set, and 82% Sensitivity and 66% Specificity on the testing set, it is demonstrated that the developed neural network can successfully recognize new patterns of the new testing group of students and effectively identify students who are at risk of making unsatisfactory performance.**

Keywords*—*machine learning, artificial neural network, higher education, student's academic progression, at-risk students

I. INTRODUCTION

The rate of students' retention, progression and completion of their educational programs not only reflects the education quality but also impacts the reputation and financial sustainability of each higher education provider. Student retention has been recognised as one of the most challenging problems for education institutions, especially for private higher education providers. In Australia, data from 2005 to 2014 shows that the attrition rate for public universities is around 15% [1] [2]. In the meantime, attrition rates for non-university higher education providers are generally higher than those of public universities, with medium attrition rate in 2014 of 25% and upper quartile attrition rates in 2014 of 32% [1] [2]. With the profound impact of the COVID-19 pandemic, student retention becomes more and more essential to ensure the quality and financial sustainability of higher education institutions.

Monitoring students' academic performance and early identifying students at risk of making unsatisfactory progression play an essential role in student retention strategies of all higher education providers. An early prediction of students' unsatisfactory performance helps the institution provide timely academic support as well as counselling to increase students' success rate. Despite the importance, most institutions find that it is a practical challenge to develop an effective strategy and framework for monitoring student's progression and identifying at-risk students. In fact, a variety of frameworks for identifying at-risk students in early stage have been developed and applied in institutions, but the complexity of indicators, the administrative burden of data analysis and the inconsistency of interpretation are undeniable. That is the reason why the application of Artificial Intelligence (AI) and Machine Learning (ML) to develop a precise, consistent and logical algorithm that can effectively predict academic performance of commencing students has recently attracted a lot of research [3] [4] [5].

Artificial neural networks have been popularly employed as a powerful tool of classification and recognition in various real-life areas which include education [6] [7] [8]. It has been widely acknowledged that the neural network can effectively model non-linear relationships between inputs and outputs, which can learn and adapt itself to new patterns and successfully solve for predicting or forecasting problems. One of the most popular training techniques is Levenberg-Marquardt (LM) algorithm which is based on the second order gradient information of an error function in order to direct the training process to a local optimal [9]. Genetic algorithm (GA) is a derivative-free global search optimization which is inspired by the natural evolution. This technique has been applied widely in evolving neural network models which can efficiently drive the training process to the global optimal [10]. The combination of GA and LM algorithm has been shown as an effective method of training neural network which can overcome the inherent drawbacks of each algorithm to ensure the training process can converge into an optimal that can produce the best classification performance [11] [12] [13].

The main objective of this paper is to propose an innovative method of early predicting academic performance of commencing students in their first semester by using artificial neural network. The combination of GA and LM algorithm will be applied to train the developed neural

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00199

network. The paper is divided into four sections. Section II provides an overview of the methodology used in our study. Results of the study will be mentioned in Section III. Section IV presents conclusions for the current study and some recommendations for future research.

II. Methods

A. Data Preparation

The data set used in this study was selected from two student cohorts commencing in Semester 1 and Semester 2 in 2023 at Sydney Institute of Higher Education (SIHE). All data are extracted anonymously to ensure research ethics and privacy.

To ensure the equality in the level of study, data from two Bachelor programs are collected, including Bachelor of Information Technology (BIT) and Bachelor of Business (BBUS). Additionally, only data from students enrolling in the full-time study mode, which means they are enrolled in four subjects, are acquired. Accordingly, an initial data set which consists of data from 359 students are collected, including:

- Students' demographic factors: Gender, Nationality
- Students' pre-enrolled academic factors: Grade 12 Results, English Entrance Test Results
- Students' post-enrolled academic factors: enrolled programs, attendance percentage in the first four weeks of students' first semester and final results of the semester.

The data are subsequently analyzed under descriptive statistics techniques to remove duplicates, outliers and defective data points (missing values). As a result, a cleaned data set of 300 students (300 data points) have been selected as the final data set to be used in the study. Each data point consists of 7 parameters including 6 inputs (defined in Table I) and 1 output. The output of each data point is the final result of the semester of each student. With the purpose of classification using neural network, the final results of students are categorized into Satisfactory (students who pass 3 or 4 out of 4 enrolled subjects in their first semester) and Unsatisfactory (students who fail 2 or more than 2 out of 4 enrolled subjects in their first semester). Accordingly, the entire data set will be divided into 2 groups: Satisfactory (which consists of 95 data points) and Unsatisfactory (which consists of 205 data points).

TABLE I. Definition of Data Inputs

	Input Name	Data Type
Input 1	Gender	Categorical
Input 2	Nationality	Categorical
Input 3	Grade 12 Results	Numerical
Input 4	English Entrance Test Results	Numerical
Input 5	Enrolled Program	Categorical
Input 6	Attendance Percentage in the first 4 weeks	Numerical

To prepare data for the classification, a step of data pre-processing will be implemented so that inputs will be suitable to be fed into the neural network. To do this, all Categorical inputs will be converted into numbers, and all Numerical inputs will be normalized. Details of this data pre-processing step is shown as follows:

- Input 1: Students' gender can take 2 values of Female and Male which are converted into 0 (Female) and 1 (Male)
- Input 2: Student's nationalities can take a variety of values. Nevertheless, based on the fact that the majority of students at SIHE are from subcontinental countries, the input 2 of the data set are converted into: 3 (India), 2 (Pakistan), 1 (Nepalese) and 0 (all other nationalities).
- Input 3: Because of the difference in grading system between different countries, each student's academic result in Grade 12 is converted to an equivalent percentage.
- Input 4: Within the selected data set of 300 students used in this study, 143 students used the International English Language Testing System (IELTS) results and 157 students used the Pearson Test of English (PTE) results to apply to SIHE. Because of the difference in grading system between the two tests, the IELTS scores of all students are converted to equivalent PTE scores.
- Input 5: Students' enrolled programs can take 2 values BIT and BBUS which are converted into 0 (BBUS) and 1 (BIT).
- Input 6: At SIHE, the average percentage of attendance of all enrolled subjects in the first four week of each semester is calculated and used as a trigger for identifying students who are at risk of making unsatisfactory progression in that semester.

B. Classification using Neural Network

In this study, a feed-forward three-layer neural network is developed as a classification unit. The structure of the neural network is shown in Fig. 1.

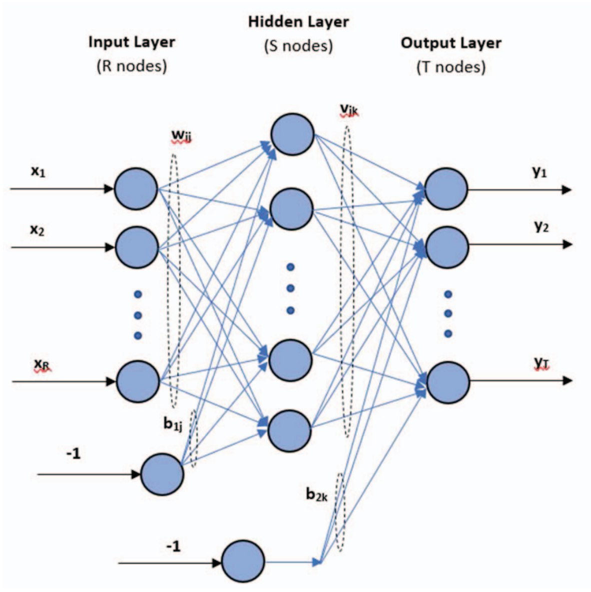

Fig. 1. Neural Network Structure

The input-output relationship of the developed neural network can be expressed as follows:

$$y_k = \sum_{j=1}^{S}\left(v_{jk} tansig\left[\sum_{i=1}^{R}(w_{ij}x_i - b_{1j})\right]\right) - b_{2k}$$

where $w_{ij}, i = 1,2,\ldots,R, j = 1,2,\ldots,S$, is the weight of the link between i-th input node and the j-th hidden node; $v_{jk}, j = 1,2,\ldots,S, k = 1,2,\ldots,T$, is the weight of the link between j-th hidden node and the k-th output node; b_{1j} , $b_{2\mathrm{k}}$ are the biases for the input layer and hidden layer respectively; R is the number of hidden nodes; S is the number of inputs; T is the number of outputs; *tansig* is the hyperbolic tangent sigmoid transfer function of the hidden layer:

$$tansig(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$$

In developing neural network, the training algorithm plays the most important role in order to achieve a desired classification performance. In this study, a 2-step training process is applied as shown in Fig. 2. The process includes a step of Global search and a step of Local search which combines both advantages of GA and LM in training neural network. The error function used for training is defined as the mean squared error (mse) of the output and its corresponding target. The trials and errors method will be applied to determine the number of hidden nodes which gives the best classification performance.

In brief, GA is employed to evolve neural network parameters by searching over the whole domain and direct the training process to the global optimal region. To do this, a population of chromosomes or individuals is initialized. Each chromosome is expressed by $[w_{ij}, v_{jk}, b_{1j}, b_{2k}]$ which means that the length of chromosome is equal to the number of neural network parameters. During the evolution, a fitness function on each chromosome is estimated which is defined as follows:

$$f(\text{chromosome}) = \frac{1}{1 + mse}$$

The population is evolved by GA until the terminating condition is fulfilled. The best chromosome with the highest value of fitness function in the last updated population is considered as the final solution of the GA algorithm or the final set of network parameters produced by the Global Search step. Then, the second step of Local Search will be implemented, using the parameters set obtained by the GA algorithm as the initial parameters for the developed neural network. The LM algorithm is then employed on this parameter set to continue training the neural network. In this way, the LM algorithm acts as a fine tuner to help the training process quickly converge toward the local solution. At each iteration of LM training, the cross-validation technique will be applied to avoid bad generalization due to overtraining. In order to do this, the training data set is divided into a training subset and a validation subset. During the training process, the error function will be monitored on both subsets. The validation error normally decreases during the initial phase of training, so does the training error. However, when the network begins to over-fit the data, the error on the validation set typically begins to increase while the training error continues to decrease. When the validation error keeps increasing for a given number of iterations, the training is stopped. The network parameters at that stopped iteration will be used as the final neural network weights and biases.

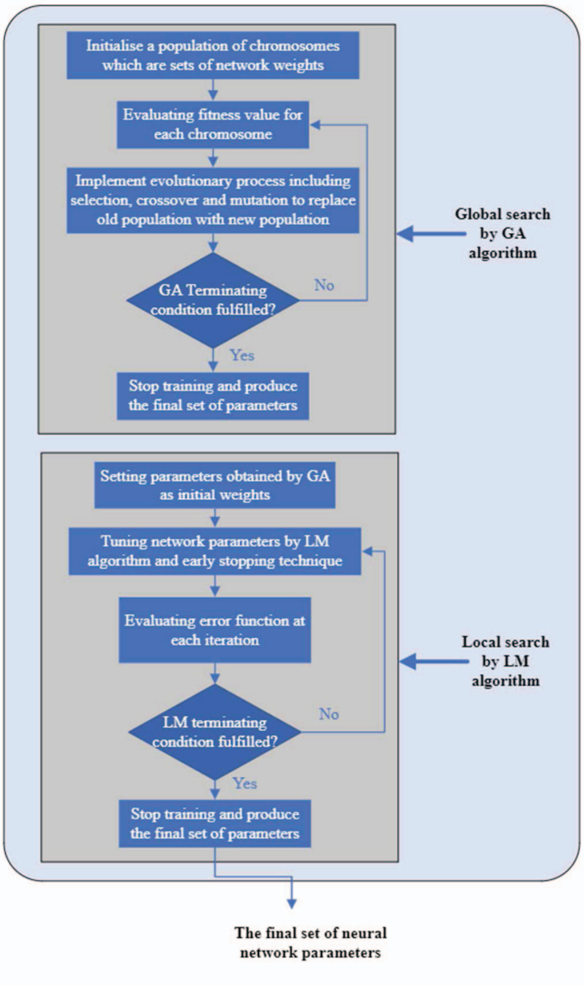
Initialise a population of chromosomes which are sets of network weights
Evaluating fitness value for each chromosome
Implement evolutionary process including selection, crossover and mutation to replace old population with new population
GA Terminating condition fulfilled?
No
Yes
Stop training and produce the final set of parameters
Global search by GA algorithm
Setting parameters obtained by GA as initial weights
Tuning network parameters by LM algorithm and early stopping technique
Evaluating error function at each iteration
LM terminating condition fulfilled?
No
Yes
Stop training and produce the final set of parameters
Local search by LM algorithm
The final set of neural network parameters

Fig. 2. GA-LM training process

C. *Result Interpretation and Evaluation Method*

In this study, to evaluate the accuracy of the proposed method, after determining the final parameters for the neural network, the classification performance of the final network will be estimated by evaluating the Sensitivity and Specificity of the classification on each data set. Specifically, for the application of predicting At-Risk students, two criteria of Sensitivity and Specificity are defined as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

where:

- True Positive (TP) is the number of students who are predicted as At-Risk students and make unsatisfactory progression at the end of the semester.
- True Negative (TN) is the number of students that are predicted as normal students and actually make satisfactory progression at the end of the semester.

- False Positive (FP) is the number of students who are predicted as At-Risk students but actually successfully progress at the end of the semester.
- False Negative (FN) is the number of students that are predicted as normal students and but actually make unsatisfactory progression at the end of the semester.

It is explicit that for any classification or prediction problems, there is always a tradeoff between the true positive rate (Sensitivity) versus the false positive rate (1-Specificity). This tradeoff can be managed by adjusting the threshold of the neural network's output. To accomplish this, a Receiver Operation Characteristic (ROC) curve will be plotted for the training set to find a suitable output threshold to improve the prediction performance. It is noted that for the application of predicting student performance, the Sensitivity, which represents the developed network's capability to correctly predict At-Risk students, is more prioritized than the Specificity. Hence, in this research, the output threshold will be selected at the point producing prediction Sensitivity of 85%, which might lead to a lower but still reasonable Specificity.

III. RESULTS

A. *Data Analysis Results*

In this research, data from 300 students are treated as 300 anonymous data points. The total data set is separated into two groups: the Satisfactory group which includes 95 data points from students who successfully progressed in their first semester and the Unsatisfactory group which includes 205 data points from students who made unsatisfactory progression in their first semester. Each data point consists of six demographic and academic parameters as listed in Table I. Hypothesis testing is applied to compare and determine the significance of differences between the two data groups as showed in Table II. In this study, chi-squared test is conducted for categorical parameters and *t*-test is conducted for numerical parameters. In all tests, *p*-values less than 0.05 are considered to be significant. Significant tests are reported in bold in Table II.

TABLE II. STATISTICAL RESULTS

Parameters	**Group comparison *p*-values**
Gender	**< 0.05**
Nationality	0.057
Grade 12 Results	0.064
English Entrance Test Results	0.123
Enrolled Program	**< 0.05**
Attendance Percentage in the first 4 weeks	**< 0.001**

Attendance percentage is shown as the strongest parameter (**p<0.001**) which significantly indicates the differences between the two groups of Satisfactory and Unsatisfactory students. This has been broadly recognized by higher education institutions to be one of the most important factors contributing to the completion rate in the first semester of commencing students who are typically experiencing critical transitional phases, such as transition in studying environment from high school to higher education, transition in culture from their home country to a new country, etc. This further affirms the importance of implementing student support strategies to help them overcome difficulties in their new studying period and maintain their commitment of participation in classes.

Statistical results also reveal slightly significant differences in Gender and Enrolled Programs between the two groups of students. It has been acknowledged that Gender is an effective attribute which is commonly used in predicting students' performance [14]. This can be explained by the remarkable differences between learning habits, motivation and studying strategies between male and female students. In addition, it is noted that other demographic attributes like students' age, family background, parents' education history are widely used in other research [14], however, due to the burden of data processing, these attributes will not be considered in this study.

The differences in Nationality, Grade 12 Results and English Entrance Test Results are recognized as insignificant. These results are predictable because all students must meet admissions criteria, including English Language Proficiency requirements and Secondary Education Results requirements. Additionally, the conversion between different grading systems is likely to have a negative impact on the accuracy of the comparison analysis. Even though these parameters are insignificant, they are still used as inputs to neural network. The reason is mainly due to the fact that the developed neural network is strongly believed to have the capability to recognize and model the non-linear relationship between its inputs and output.

B. *Classification Results*

The structure of the developed neural network encompasses six input nodes which are six parameters extracted from the data-preprocessing steps as mentioned in section II. The neural network has one output node that represents the status of students at the end of their first semester as Satisfactory or Unsatisfactory. The desired output is set at 1 in the case of Unsatisfactory and -1 in the case of Satisfactory. The output cutoff threshold is determined as the point producing 85% Sensitivity on the ROC curve. The number of hidden nodes is varied from 5 to 15. The final number of hidden nodes is selected as the one that yields the best prediction performance. Resultantly, it is determined that with 6 input nodes and 1 output node, the best prediction results are produced by the network with 11 nodes in the hidden layer.

For training the developed neural network, the overall data set is randomly divided into a training set and a testing set as follows:

- The training set consists of 200 data points including 140 Unsatisfactory points and 60 Satisfactory points.
- The testing set consists of 100 data points including 65 Unsatisfactory points and 35 Satisfactory points.

For LM training step, to implement cross-validation, the training set is subsequently subdivided into two subsets, a LM-training subset and a LM-validation subset with a ratio of 2:1 as follows:

- The LM-training subset consists of 135 data points including 95 Unsatisfactory points and 40 Satisfactory points.

- The LM-validation subset consists of 65 data points including 45 Unsatisfactory points and 20 Satisfactory points.

TABLE III. CLASSIFICATION RESULTS

	Training set		Testing set	
	Sen[a]	**Spe**[b]	**Sen**[a]	**Spe**[b]
Mean Performance	85%	63%	79%	64%
Best Performace	85%	69%	82%	66%

[a] Sen: Sensitivity; [b] Spe: Specificity

Classification results are shown in Table III. The reported results are the mean performance and best performance of 20 running times. With the best performance of 85% Sensitivity and 69% Specificity on the training set, and 82% Sensitivity and 66% Specificity on the testing set, it is demonstrated that the developed neural network has good generalization which indicates that the trained network can recognize the new patterns of new students and successfully predict students' academic progression at the end of their first semester.

It is noted that in this study, the training set and the testing set are generated by randomly dividing data from the mixed pool of students commencing in Semester 1 and Semester 2 of the Academic Year 2023 at SIHE. This can be considered as a limitation of the current study as the trained neural network should be ideally tested with data from an unseen cohort of students. This is mainly due to the fact that each cohort of students commencing in each semester are likely to be subject to the changes in admissions requirements and marketing strategies of the institution. Nevertheless, due to the fact that SIHE is a newly established institution in late 2022, the shortage of data due to the limited semesters of delivery is understandable and unavoidable.

For comparison and analysis purpose, two other prediction methods including Multiple Linear Regression and Decision Tree are implemented. These are two statistics-based methods which are widely used to predict students' performance [15] [16]. Classification results of each method on the same testing set are presented in Table IV.

TABLE IV. COMPARISON BETWEEN METHODS

Method	**Sensitivity**	**Specificity**
Multiple Linear Regression (MLR)	61%	59%
Decision Tree (DT)	73%	68%
GA+LM Neural Network	**82%**	**66%**

The results indicate that the developed neural network achieves better classification results compared to the other two methods. It is obvious the results obtained by MLR is incompetent with only 61% Sensitivity and 59% Specificity. These results are foreseeable due to MLR's inherent drawbacks of overfitting, especially in the application with large sample size where the regression model represents the data noise rather than the real relationships between variables. On the other hand, the results obtained by Decision Tree are competitive with 73% Sensitivity and 68% Specificity. However, the instability to changes in data patterns of this method makes it less suitable for the application of predicting student performance in this study. The results comparison and analysis show that the proposed AI-based method of using neural network outperforms the statistics-based methods in the application of predicting student progression and identifying students at risk of making unsatisfactory progression.

IV. CONCLUSIONS

This paper presented a method of predicting students' academic progression by using neural network for the purpose of early identifying students at risk of making unsatisfactory progression at the end of their first semester. Using anonymous data from 300 students at Sydney Institute of Higher Education (SIHE), six parameters from students' profile were extracted, processed and analyzed. Statistical results indicated that students' attendance percentage in the first four weeks of the semester is the most significant indicator for identifying at-risk students. All six parameters were then used as inputs of a neural network classification unit. The network was trained by a 2-step GA+LM strategy which combines the GA's global search capability and LM's local search capability. The achieved classification results indicated that the developed neural network can effectively predict students' academic performance and identify students at risk of making unsatisfactory progression. This early identification of at-risk students plays an important role in enhancing students' success rate by offering timely and effective academic and non-academic support services to help those students back on track of normal studying pattern and performance.

For future research, more advanced algorithms are suggested to be explored to enhance the generalization and overall classification performance of the developed neural network. The limitation of data shortage in creating training and testing sets will be addressed and overcome in the future when new cohorts of students commencing their studies at SIHE. Additionally, different approaches will be examined to investigate more students' attributes, especially academic attributes after the commencement of studying, e.g. the number of hours students spend on online Learning Management System in their first four weeks of each semester or students' engagement in student support activities, etc.

REFERENCES

[1] "Final Report - Improving retention, completion and success in higher education," Higher Education Standards , 2017.

[2] "Characteristics of Australian higher education providers and their relation to first-year student attrition," TEQSA, June 2017.

[3] S. Sarwat, N. Ullah, S. Sadiq, R. Saleem, M. Umer, A. A. Eshmawi, A. Mohamed, and I. Ashraf, "Predicting Students' Academic Performance with Conditional Generative Adversarial Network and Deep SVM," Sensors, 22, 4834, 2022.

[4] W. Villegas-Ch, J. Govea, and S. Revelo-Tapia, "Improving Student Retention in Institutions of Higher Education through Machine Learning: A Sustainable Approach," Sustainability, 2023.

[5] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms" Smart Learn. Environ. 9, 11, 2022.

[6] Baashar, Yahia, G. Alkawsi, A. Mustafa, A. A. Alkahtani, Y. A. Alsariera, A. Q. Ali, W. Hashim, and S. K. Tiong, "Toward Predicting Student's Academic Performance Using Artificial Neural Networks (ANNs)," Applied Sciences 12, no. 3: 1289, 2022.

[7] C. F. Rodríguez-Hernández, M. Musso, E. Kyndt, and E. Cascallar, "Artificial neural networks in academic performance prediction: Systematic implementation and predictor evaluation, " Computers and Education: Artificial Intelligence, Volume 2, 2021.

[8] E. Okewu, P. Adewole, S. Misra, R. Maskeliunas and R. Damaseviciu s, “Artificial Neural Networks for Educational Data Mining in Higher Education: A Systematic Literature Review,” Applied Artificial Intelligence, 35:13, 983-1021, 2021.

[9] M. T. Hagan and M. B. Menhaj, “Training feedforward networks with the Marquardt algorithm,” IEEE Transactions on Neural Networks, vol. 5, pp. 989-993, 1994.

[10] D. J. Montana and L. Davis, “Training feedforward neural networks using genetic algorithms,” Proceedings of the 11th international joint conference on Artificial intelligence, Volume 1, Detroit, Michigan: Morgan Kaufmann Publishers Inc., 1989.

[11] L. B. Nguyen, A. V. Nguyen, S. H. Ling and H. T. Nguyen, “Combining genetic algorithm and Levenberg-Marquardt algorithm in training neural network for hypoglycemia detection using EEG signals,” 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, pp. 5386-5389, 2013.

[12] H. Mahmoudabadi, M. Izadi and M. B. Menhaj, “A hybrid method for grade estimation using genetic algorithm and neural networks,” Comput Geosci 13, 91–101, 2009.

[13] Y.R. Ding, Y.J. Cai, P.D. Sun and B. Chen, “The Use of Combined Neural Networks and Genetic Algorithms for Prediction of River Water Quality,” Journal of Applied Research and Technology, Volume 12, Issue 3, Pages 493-499, 2014.

[14] A. M. Shahiri, W. Husain, N. A. Rashid, “A Review on Predicting Student's Performance Using Data Mining Techniques, Procedia Computer Science,” Volume 72, Pages 414-422, 2015.

[15] A. Kumar, K. K. Eldhose, R. Sridharan and V. V. Panicker, “Students' Academic Performance Prediction using Regression: A Case Study,” 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India, pp. 1-6, 2020.

[16] H. Hamsa, S. Indiradevi and J. J. Kizhakkethottam, “Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm,” Procedia Technology, Volume 25, Pages 326-332, 2016.