J Inf Process Syst, Vol.20, No.2, pp.263~272, April 2024
https://doi.org/10.3745/JIPS.02.0213

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

An Early Warning Model for Student Status Based on Genetic Algorithm-Optimized Radial Basis Kernel Support Vector Machine

Hui Li[1,2,*], Qixuan Huang[1,2], and Chao Wang[3]

Abstract

A model based on genetic algorithm optimization, GA-SVM, is proposed to warn university students of their status. This model improves the predictive effect of support vector machines. The genetic optimization algorithm is used to train the hyperparameters and adjust the kernel parameters, kernel penalty factor C, and gamma to optimize the support vector machine model, which can rapidly achieve convergence to obtain the optimal solution. The experimental model was trained on open-source datasets and validated through comparisons with random forest, backpropagation neural network, and GA-SVM models. The test results show that the genetic algorithm-optimized radial basis kernel support vector machine model GA-SVM can obtain higher accuracy rates when used for early warning in university learning.

Keywords

Early Warning Model, Genetic Algorithm Optimization, Radial Basis Kernel, Support Vector Machine

1. Introduction

Many university students struggle with academic pressure, interpersonal relationships, and uncertainty about their future, which are manifested as mental health problems, study anxiety, and poor study habits, among others. These issues affect the growth and development of the student [1, 2].

There is a trend of using machine-learning algorithms for analysis and prediction in the context of early warnings for students in higher education. In addition to the support vector machine (SVM) algorithm, other machine learning algorithms such as random forest and backpropagation neural networks have been applied in this field [2–5]. However, in practical applications, the results of different algorithms may vary, and there are some limitations in existing research [6, 7].

The motivation in this study was to address the problems faced by college students, whether in terms of academic, mental health, or social issues [8], by using a combination of SVM and genetic optimization algorithms to improve the accuracy of student warnings. Unlike previous studies, this method combines multiple results to predict the status of students more comprehensively. The unique feature of this study is that it not only focuses on academic and mental health issues but also on social issues, to help better support the students' development.

Manuscript received July 11, 2023; first revision November 6, 2023; second revision December 7, 2023; accepted December 9, 2023.

* **Corresponding Author:** Hui Li (hdcll2008@163.com)
[1] School of Information Engineering, Handan University, Handan, China (hdcll2008@163.com, tma3919@163.com)
[2] Hebei key Laboratory of Optical Fiber Biosensing and Communication Devices (SZX2022010), Handan, China (hdcll2008@163.com, tma3919@163.com)
[3] School of Software, Handan University, Handan, China (rjxyky@163.com)

2. Related Studies

2.1 Support Vector Machine Principle

In this study, we chose the most widely used radial basis kernel function, which is expressed as

$$k(x_i, x_j) = e^{-g||x_i - x_j||}, \tag{1}$$

where g is the kernel parameter, representing the width of the radial basis kernel function action, and C is the penalty factor. Therefore, to improve the accuracy of the machine learning warning model in this study, the parameters should be chosen appropriately.

2.2 Support Vector Machine Optimization Model based on Genetic Algorithm

In this study, genetic algorithms were used to select suitable parameters, whereby the parameter selection process was optimized to build a more accurate machine learning warning classification model. By combining traditional SVMs and genetic algorithms, this study proposes an early warning learning model based on improved SVMs, with objective function as follows:

$$F = minf(z_1, \ldots, z_i) = min\frac{1}{l}\sum_{i=1}^{l}\left(y_i - f(x_i)\right)^2, \tag{2}$$

where l represents the number of samples, y_i represents the actual value, and $f(x_i)$ represents the predicted value.

2.3 Existing Research

According to existing studies [9, 10], supervised learning is the most commonly used data mining technique to solve problems related to the classification of mental health problems. The most commonly used algorithms include SVMs, followed by decision tree and neural networks. All three models have a high degree of accuracy, which is in excess of 70%, and good generalization ability, which can prevent overfitting.

After testing, genetic algorithm-optimized support vector machine (GA-SVM) performed the best in the training and testing scenarios on the same dataset. The results showed that genetic algorithms can effectively search the hyperparameter space of an SVM and determine the optimal hyperparameter configuration, thereby improving the performance of the classifier. In addition, the random nature of genetic algorithms enables them to escape local optima, making them more likely to find the global optimum and enhance the generalization ability of the model

3. GA-SVM Early Warning Forecasting for Students in Higher Education

Given the abstract, nonlinear, and categorical nature of the student warning problem and small sample size, a SVM algorithm was used to solve the classification problem. SVMs have many advantages, such as neither being affected by sample size nor being prone to overfitting. Therefore, this study used a SVM algorithm for model training.

3.1 Characteristics of Early Warning Models for Students in Higher Education

This study relied on publicly available open-source datasets of student mental health, and a combination of extensive literature research, student interviews and teacher recommendations was used to screen for the characteristics of learning crises. After grouping these characteristics into mental health, academic, and social components, the characteristics of these three components were combined.

In accordance with the principle of "considering the causes, capturing the key elements, reducing the cost of prediction, and facilitating problem solving," each indicator element was refined to an easily measurable level. This downgrading process made the final indicator element more operational, laying the foundation for subsequent information collection. After several screenings, 16 key characteristics of crisis generation that affect learning were extracted, as shown in Table 1. The data for these indicators can be easily obtained from open-source datasets.

Table 1. Features of a student warning model

Serial number	Parameters
1	Age
2	Major
3	Gender
4	Cumulative grade point average (CGPA)
5	Sleep quality
6	Physical exercise
7	Diet quality
8	Social support
9	Emotional relationship status
10	Counseling service utilization
11	Family history
12	Chronic illness
13	Financial stress
14	Extracurricular activity participation
15	Semester credit load
16	Accommodation type

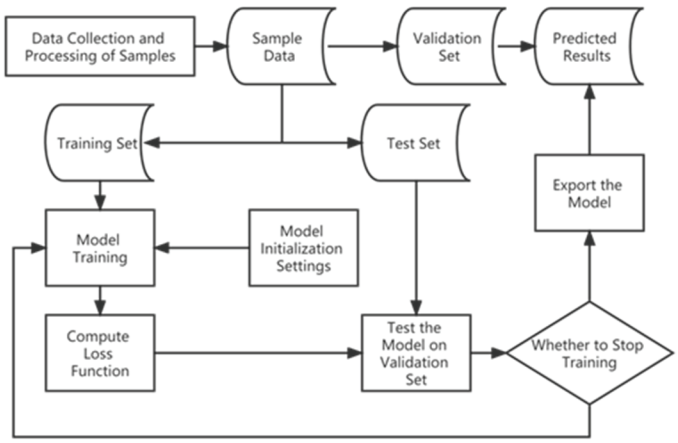

Fig. 1. Implementation flowchart of GA-SVM student warning prediction model.

3.2 Learning Early Warning Model based on GA-SVM

According to actual needs, the input and output variables of the model (the presence of student warning) were determined, and the data of 4,700 current junior students were exported through each system, processed, and merged. The optimal hyperparameters were confirmed using a genetic algorithm, and the SVMs was trained using the results of the best search. The model was trained on the training set; the reliability of the model was verified using the validation set; and finally, 20% of the test set was used for comparisons with other models. The implementation of the GA-SVM predictive student-warning model is shown in Fig. 1.

3.2.1 Data selection and pre-processing

Input and output variables: The 16 impact factors mentioned above were chosen as input variables, with yes and no constituting the output variables. Yes means that the student has an early warning situation; that is, they may have psychological, academic, or social problems. No implies that there are no such problems. In the dataset, data that lacked factors, such as null sleep quality, were deleted. After data removal, the cumulative grade point average (CGPA) parameters were classified as follows: CGPA greater than 3.0 as A; CGPA less than 3.0 but greater than 2.5 as B; CGPA greater than 2.0 but less than 2.5 as C; and CGPA less than 2.0 as D. In this way, the characteristics were converted into numeric types. Finally, all datasets were subjected to one-hot encoding, which converted all categorical variables into vector form.

Training and test data selection: To run the model more easily and ensure its accuracy, the samples were first screened to remove those with unusable or missing data. Subsequently, through cross-validation, 80% of the data were used as the training set, of which 20% were selected as the validation set, and the remaining 20% were used as the test set, as shown in Fig. 2. This helped improve the reliability of the model.

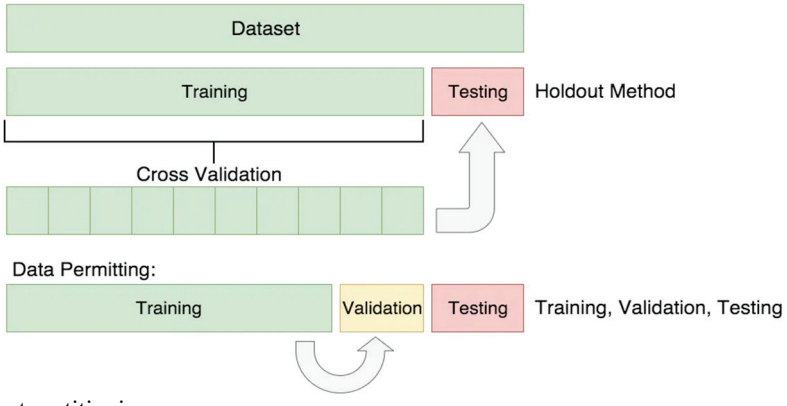

Fig. 2. Data-set partitioning.

3.2.2 Model parameter settings

Fitness function: In our implementation of the genetic algorithm, the fitness function was used to evaluate the performance of each individual. Specifically, we employed a fitness function based on classification accuracy. This implied that the fitness of each individual is calculated based on its accuracy in the classification task. The formula for the fitness function is expressed as follows:

$$Fitness = \frac{Number\ of\ Correctly\ Classified\ Samples}{Total\ Number\ of\ Samples}. \quad (3)$$

This ensures that individuals with higher fitness levels perform better upon completing the classification task.

Encoding: In our genetic algorithm, each individual was encoded using a real-number encoding method. This means that the chromosome of each individual consists of a sequence of real numbers, each representing different hyperparameters, such as the penalty parameter C and the kernel function parameter γ of SVM. This encoding method not only enhances the precision of parameter search but also makes the algorithm more flexible and efficient.

Parameters of the crossover operation: This study adopted a uniform crossover strategy to perform crossover operations. In this process, two parent individuals are selected and genes from the parent individuals are randomly exchanged based on a specified crossover probability, thus generating new offspring individuals. We set the crossover probability to 0.6, meaning that there was a 60% chance of selecting genes from one parent and a 40% chance from the other parent. This method has the advantage of maintaining diversity within the population, while also promoting the transfer of useful traits.

3.2.3 Model prediction performance evaluation criteria

The performance of the model was evaluated using accuracy, precision, recall, and F1-score, which better reflect the performance of the classification model as they take into account the classification accuracy of the model, as well as false positives and false negatives.

4. Experimental Cases

4.1 Analysis of Forecast Results

In this study, Python was used to optimize the SVM with the genetic algorithm, running it several times to obtain the best parameters after each iteration. This process is illustrated in Fig. 3.

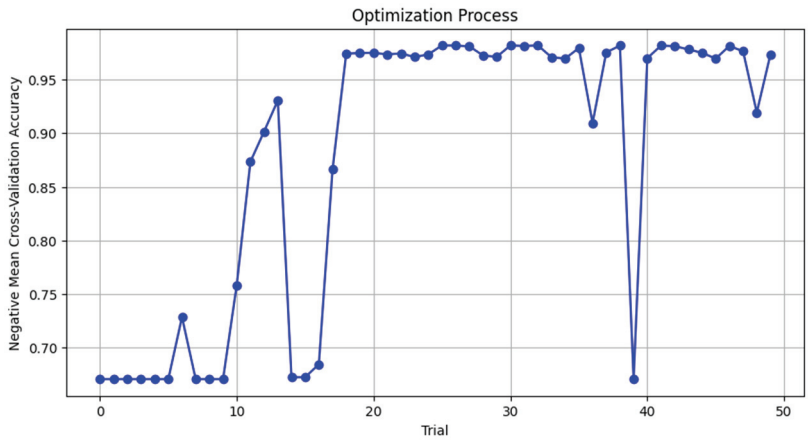

Fig. 3. Fitness curve plot for genetic algorithm optimization.

After training on the training set and validating on the validation set, the expected results were obtained. Finally, testing on the test set yielded an accuracy of 97.29%. The other parameters such as precision, recall, and F1-score were 0.9630, 0.9879, and 0.9801, respectively. These four parameters were used for the comparisons, to validate the prediction results.

For dichotomous classification problems, the confusion matrix provides a more intuitive picture of the model classification. In the confusion matrix diagram, true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are represented by different colors or patterns. Through confusion matrix plots, the model performance in each category can be visualized in relation to error type. The plot of predicted sample and true value comparison provides a visual representation of the prediction. The results are shown in Figs. 4 and 5.

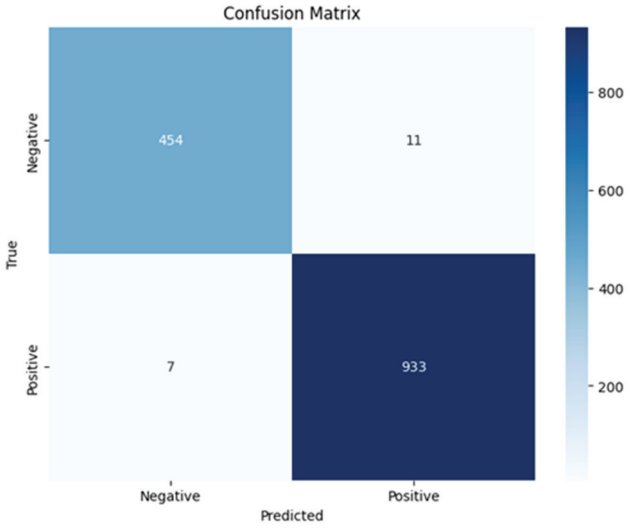

Fig. 4. GA-SVM confusion matrix plot.

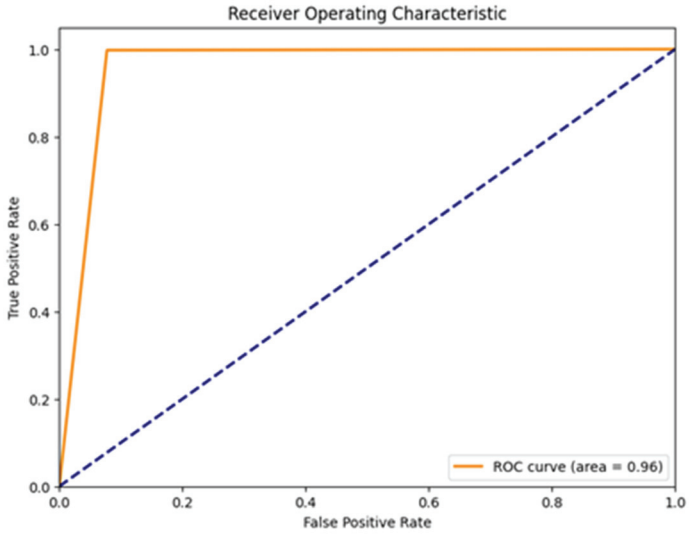

Fig. 5. GA-SVM receiver operating characteristic.

4.2 Verifying the Validity of the Prediction Results

For a fair and effective comparison, this study employed commonly used hyperparameter settings to train the random forest algorithm, multilayer perceptron (MLP), extreme gradient boosting (XGBoost), decision tree, and k-nearest neighbors (KNN) models. This means that the hyperparameters of all the models were not specifically optimized, thereby ensuring that the comparison results more accurately reflect the performance of each model under standard configurations. This approach not only ensures the simplicity and reproducibility of the experimental design but also provides a balanced benchmark for assessing the effectiveness of the SVM learning early warning model optimized using a genetic algorithm, in comparisons with other standard machine learning methods. In this study, these models were

Table 2. Performance of different models

Model	Accuracy	Precision	Recall	F1-score
MLP	0.9081	0.8843	0.9148	0.9353
Random forest	0.8960	0.9250	0.9191	0.9220
XGBoost	0.9572	0.9924	0.9372	0.9670
Decision tree	0.9067	0.9153	0.9443	0.9296
KNN	0.9601	0.7852	0.8700	0.8254
GA-SVM	0.9729	0.9630	0.9879	0.9801

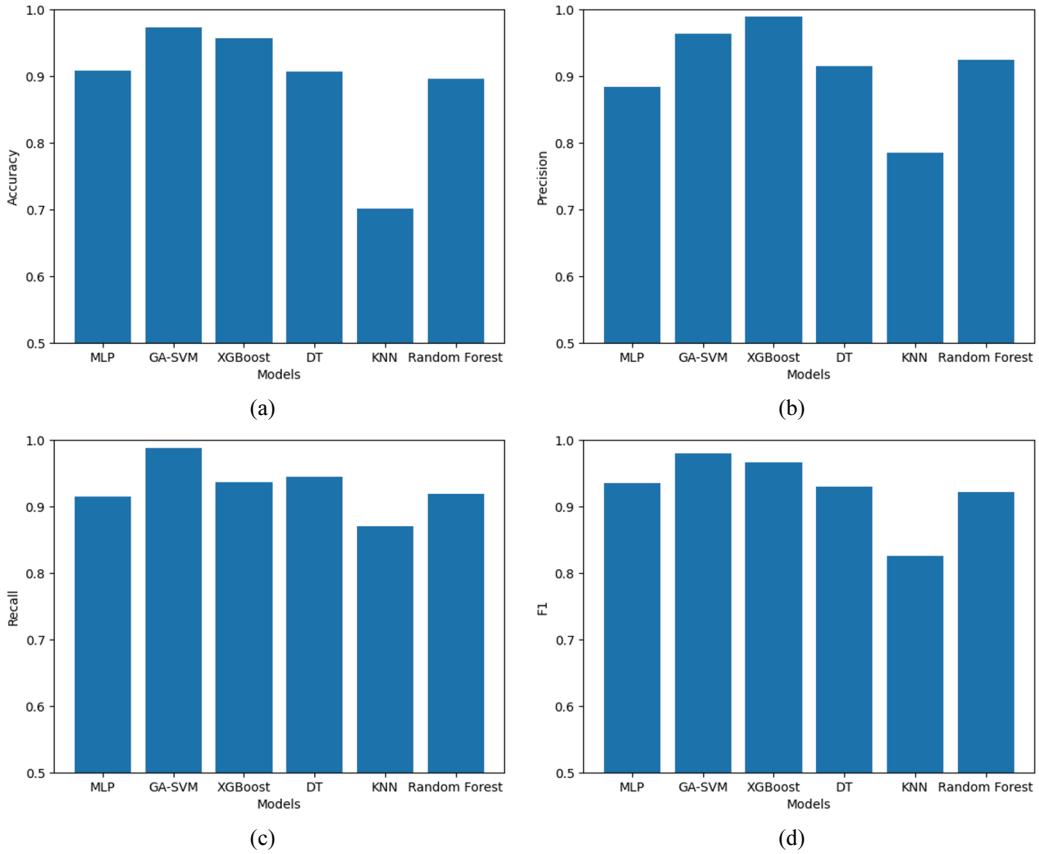

Fig. 6. Comparison of different models: (a) accuracy, (b) precision, (c) recall, and (d) F1-score.

trained on the same data samples, using the same test set for prediction, to compare based on accuracy, precision, recall, and F1-score. The score results of all models are shown in Table 2. Model parameters are compared in Figs. 6 and 7.

Compared to the prediction results of other models, the genetic algorithm-optimized SVM had a better predictive effect on students' learning status, followed by XGBoost, with the worst being KNN. Although there may be some uncertainty and error in the prediction accuracy under the influence of small sample data size, the prediction results of the genetic algorithm-optimized SVM can have a certain reference value in judging students' learning status, based on the comparisons in a previous article.

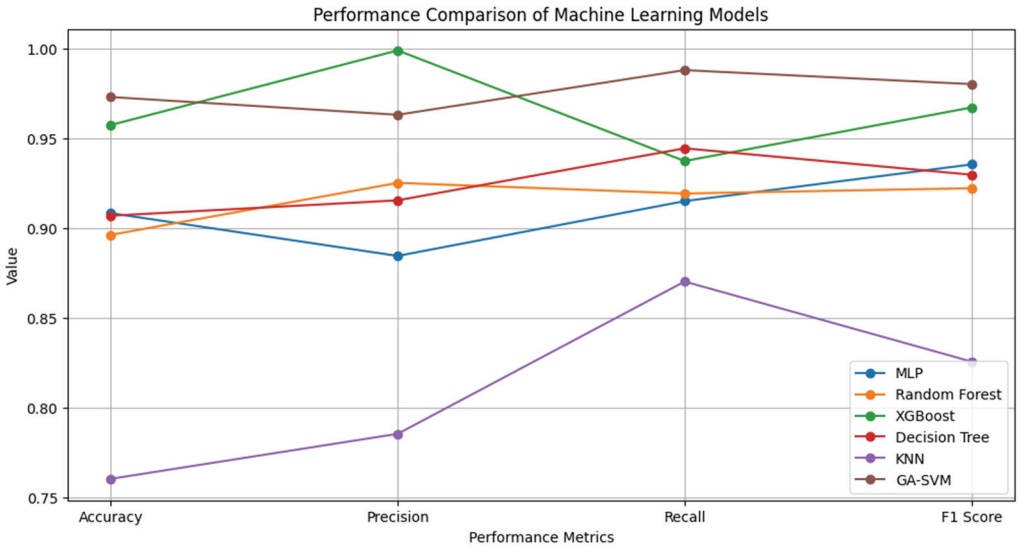

Fig. 7. Performance comparison of models.

5. Concluding Remarks

In summary:

1) Using genetic algorithms can effectively solve the hyperparameter selection problem and provide randomness.

2) The results of comparing the two different prediction models show that the SVM learning warning model based on genetic algorithm optimization has higher prediction accuracy and smaller error. Thus, it can be used to determine the learning status of students, which has certain application value in predicting student status.

3) This study further emphasizes the significance of optimization algorithms, particularly in the context of hyperparameter selection. Our findings demonstrate that the use of genetic algorithms not only addresses the challenges associated with selecting appropriate hyperparameters but also introduces an element of randomness that is crucial in navigating the complex landscape of parameter tuning. This approach has proven to be particularly effective in enhancing the performance and reliability of machine learning models.

In addition, future research may focus on the following areas.

- How to identify student problems more accurately, rather than dichotomizing to yes and no.
- How to choose more general factors for prediction so that the model can be generalized.
- How to develop more efficient optimization algorithms to reduce optimization time. Genetic algorithms require longer time to optimize, sacrificing time in exchange for performance.

Acknowledgement

This paper is funded by Science and Technology Project of Hebei Education Department (No. QN20 21405) and Handan Science and Technology Research and Development Plan Project (No. 21422021173 and 21422031170) and Research Fund of Handan University (No. XZ2021202).

References

[1] N. S. M. Shafiee and S. Mutalib, "Prediction of mental health problems among higher education student using machine learning," *International Journal of Education and Management Engineering*, vol. 10, no. 6, pp. 1-9, 2020. https://doi.org/10.5815/ijeme.2020.06.01

[2] L. Zhou and W. An, "Data classification of mental health and personality evaluation based on network deep learning," *Mobile Information Systems*, vol. 2022, article no. 9251598, 2022. https://doi.org/10.1155/2022/9251598

[3] C. Liu, H. Wang, Y. Du, and Z. Yuan, "A predictive model for student achievement using spiking neural networks based on educational data," *Applied Sciences*, vol. 12, no. 8, article no. 3841, 2022. https://doi.org/10.3390/app12083841

[4] D. Jia and H. Zhao, "Optimization of entrepreneurship education for college students based on improved random forest algorithm," *Mobile Information Systems*, vol. 2022, article no. 3682194, 2022. https://doi.org/10.1155/2022/3682194

[5] S. Chen, "Improved fuzzy algorithm for college students' academic early warning," *Mathematical Problems in Engineering*, vol. 2022, article no. 5764800, 2022. https://doi.org/10.1155/2022/5764800

[6] D. Zhu, "A fuzzy comprehensive evaluation and random forest model for financial account audit early warning," *Mobile Information Systems*, vol. 2022, article no. 5425618, 2022. https://doi.org/10.1155/2022/5425618

[7] Y. Gu, Y. Wu, H. Huang, and Q. Pang, "Prediction model of dam safety behavior based on genetic algorithm optimized support vector machine," *Journal of Hohai University (Natural Sciences)*, vol. 48, no. 5, pp. 419-425, 2020. https://doi.org/10.3876/j.issn.1000-1980.2020.05.006

[8] X. Deng, "A fuzzy qualitative simulation study of college student's mental health status," *Discrete Dynamics in Nature and Society*, vol. 2022, article no. 5177969, 2022. https://doi.org/10.1155/2022/5177969

[9] A. Priya, S. Garg, and N. P. Tigga, "Predicting anxiety, depression and stress in modern life using machine learning algorithms," *Procedia Computer Science*, vol. 167, pp. 1258-1267, 2020. https://doi.org/10.1016/j.procs.2020.03.442

[10] F. Ge, Y. Li, M. Yuan, J. Zhang, and W. Zhang, "Identifying predictors of probable posttraumatic stress disorder in children and adolescents with earthquake exposure: a longitudinal study using a machine learning approach," *Journal of Affective Disorders*, vol. 264, pp. 483-493, 2020. https://doi.org/10.1016/j.jad.2019.11.079

Hui Li **https://orcid.org/0009-0003-1529-8508**

She received M.S. degree in School of Information and Electrical Engineering from Hebei University of Engineering in 2008, P.R. China. Now, she teaches in Handan University, Associate Professor. Her research interest include artificial intelligence and robot systems.

Qixuan Huang **https://orcid.org/0009-0008-6987-7173**

He received B.S degree in School of Network Engineering from Handan University in 2022. His current research interests include deep learning and augmented reality.

Chao Wang **https://orcid.org/0009-0001-1055-1346**

He received B.S. in School of Information Technology at Hebei Normal University in 2008 and his Master's degree in School of Educational Information Technology at South China Normal University in 2012. His current research interests include the application of artificial intelligence in education and educational informatization.