Research on Prediction and Application of Physical Fitness Test Data Using Ensemble Learning Models

Yanxia Qin[1], Mincong Chen[2]
[1]Department of Physical Education, Nanjing Institute of Technology, Nanjing 211167, Jiangsu, China
[2]School of Communication and Artificial Intelligence, School of Integrated Circuits, Nanjing Institute of Technology, Nanjing 211167, Jiangsu, China
Corresponding author: Mincong Chen (e-mail: A18151005597@163.com)

ABSTRACT This study proposes an ensemble-learning-based prediction framework for physical fitness assessment using physiological and performance data acquired from conventional testing systems and wearable sensing platforms. Student attributes, including gender, height, weight, and vital capacity, are utilized as input features, while physical fitness indicators are treated as prediction targets. To improve prediction robustness and anomaly sensitivity, a two-stage ensemble architecture combining decision trees and multilayer perceptrons is developed. The decision tree provides preliminary predictions and suppresses the influence of outliers, whereas the multilayer perceptron performs nonlinear feature refinement to enhance predictive accuracy. Experimental evaluation based on 3,000 university students demonstrates that the proposed framework consistently achieves lower prediction errors than individual learning models in both 50-meter sprint and standing long jump assessments. Furthermore, anomaly analysis enables the identification of potential measurement deviations and performance-improvement opportunities. Longitudinal validation indicates that students selected through the proposed framework and subjected to personalized interventions achieve significantly greater improvements than control groups. By interpreting physical fitness assessment as a wearable sensing and data-analysis problem, the proposed approach establishes a closed-loop process linking data acquisition, anomaly detection, performance prediction, and intervention optimization. The framework provides an engineering-oriented methodology for intelligent health monitoring, signal-based performance evaluation, and personalized training support in large-scale monitoring environments.

INDEX TERMS ensemble learning, wearable sensing, anomaly detection, intelligent health monitoring

I. INTRODUCTION

As a critical indicator for assessing college students' physical fitness, the National Physical Fitness Test for university students serves not only as a vital component of educational policies but also as a cornerstone for national human resource development and individual lifelong health [1]. However, the persistent decline in college students' physical fitness levels, fluctuating test compliance rates, and inherent systemic issues in the testing framework have raised widespread concerns. With the advancement of textile science, although high-performance sportswear and wearable textile sensors have begun to be used to assist in monitoring physical status, the core challenge remains how to accurately process and predict fitness data to provide scientific guidance. According to the National Physical Fitness Standards, university students must meet specific benchmarks in events such as the 50-meter sprint, seated forward bend, standing long jump, pull-ups (male), 1-minute sit-ups (female), 1000-meter run (male), and 800-meter run (female). For instance, at a university in Nanjing, the failure rates in 2021, 2022, and 2023 were 10.35%, 9.45%, and 12.04%, respectively [2]. Current challenges in physical fitness testing include outdated testing equipment with significant measurement errors, over-reliance on manual operations, and compromised data accuracy. Even when using advanced textile-based sensors for data collection, the lack of robust computational frameworks can lead to data misinterpretation. Additionally, the lack of tailored training programs for underperforming students has resulted in limited improvements in test scores. To address these issues, this study proposes leveraging artificial intelligence (AI) technologies and an ensemble learning-based prediction model to analyze physical fitness test data [3-5]. By inputting student characteristics (e.g., gender, height, weight, and vital capacity) as features and physical test scores as outputs, the model is trained to predict performance. Subsequent outlier analysis of predictions enables the construction of personalized training plans, allowing students to efficiently improve their test scores within constrained training periods [6].

II. DATA PREPARATION

A. PHYSICAL TEST DATA IMPORT

Figure 1 displays the 50-meter sprint test results of 500 students from a university in Nanjing. The horizontal axis represents student IDs, while the vertical axis indicates the 50-meter sprint time (in seconds). As shown in the figure, the sprint times are distributed between 6 and 11 seconds. Figure 2 presents the standing long jump performance of the same cohort, with the horizontal axis denoting student IDs and the vertical axis representing jump distances (in centimeters). The jump distances range from 120 to 280 centimeters, demonstrating a broad distribution of performance levels.

B. DATA PREPROCESSING

The contemporary university student physical fitness test data include students' basic physical characteristics such as height, weight, and vital capacity, as well as specialized test items such as the 50-meter sprint, standing long jump, sit-and-reach, pull-ups (for male students), one-minute sit-ups (for female students), 800-meter run (for female students), and 1000-meter run (for male students). These tests comprehensively evaluate and improve the health status of university students. In this study, students' basic characteristics (e.g., height, weight, and vital capacity) are used as model inputs, while their performance in specialized test items serves as the outputs.

1) Data Cleaning

Missing Values: Missing test scores were imputed with the mean value of available data, while records with multiple missing entries were directly removed.

2) Outliers: The Interquartile Range (IQR) method was applied to identify outlier boundaries. Linear interpolation method was used to obtain IQR. The steps were as follows:

First, sort the n training data in ascending order, denoted as D_1, D_2 ,..., D_n. Let n_1=0.25×(n+1) and n_2=0.75×(n+1). Compute the first quartile Q_1=D[n_1]+0.75×(D[n_1+1]-D[n_1]) and the third quartile Q_3=D[n_2]+0.25×(D[n_2+1]-D[n_2]), where [n_1] and [n_2] represent the integer values of n_1 and n_2 respectively. Calculate the interquartile range as IQR =Q_3-Q_1. Finally, determine the botom and top limits of the box: Dbotom=Q_1-1.5×IQR and Dtop=Q_3+1.5×IQR.

3) Normalization: Input features were standardized using min-max normalization [7]:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

C. DATASET PARTITIONING

The training data was randomly split into a training set and a test set. The training set was further divided into a training subset and a validation subset to optimize model performance evaluation.

III. MODELS

A. MULTILAYER PERCEPTRON (MLP)

The multilayer perceptron (MLP) consists of an input layer, hidden layers, and an output layer, as illustrated in Figure 3 [8]. The input layer is responsible for receiving raw input data, the hidden layers transform the input data into feature representations, and the output layer performs regression prediction by automatically learning the relationship between inputs and outputs. These layers are fully interconnected.

The MLP updates its weight network through forward propagation of data and backward propagation of errors, exhibiting strong nonlinear fitting capabilities, flexible architecture, and good generalization performance. In this study, we use students' gender, height, weight, and vital capacity as input features and their individual test scores as outputs for regression prediction.

B. DECISION TREE

Decision tree is a supervised learning algorithm based on a tree-like structure, which performs classification or regression tasks by recursively partitioning data. It can automatically handle missing values and outliers in raw data, making it applicable in predicting college students' physical fitness test results. The implementation process is as follows [9]: The training dataset is represented as {(x_1,y_1), (x_2,y_2),(x_3,y_3), (x_4,y_4)....., (x_n,y_n)},where x_n denotes the input features and y_n the output. Here, x_i=($x_i^{(1)}$, $x_i^{(2)}$, $x_i^{(3)}$,$x_i^{(4)}$) corresponds to four input features: gender, height, weight, and vital capacity of college students, n denotes the number of samples, while y_n represents the physical test score. The decision tree partitioning principle is as follows:

For the j-th feature variable $x^{(j)}$ and a candidate split value s, two regions are defined: $R_1(j,s) = \{x \mid x^{(j)} \leq s\}$, $R_2(j,s) = \{x \mid x^{(j)} > s\}$. The optimal split is then determined by solving [10-12]:

$$\min \left[\sum_{y_i \in R_1(j,s)} (y_i - m)^2 + \sum_{y_i \in R_2(j,s)} (y_i - n)^2 \right] \tag{2}$$

Here, m represents the mean of all y_i values in region $R_1(j,s)$ and n denotes the mean of all y_i values in region $R_2(j,s)$, calculated as:

$$m = \frac{1}{N_1} \sum_{y_i \in R_1(j,s)} y_i, \qquad n = \frac{1}{N_2} \sum_{y_i \in R_2(j,s)} y_i,$$

where N_1 and N_2 are the number of samples in $R_1(j,s)$ and $R_2(j,s)$ respectively. After identifying the optimal split point (j,s), the dataset is partitioned into two sub-regions. The same splitting process is then recursively applied to each sub-region until further subdivision is no longer feasible.

C. ENSEMBLE LEARNING MODEL

Given the inherent robustness of decision tree algorithms to outliers in raw data, a two-stage ensemble framework

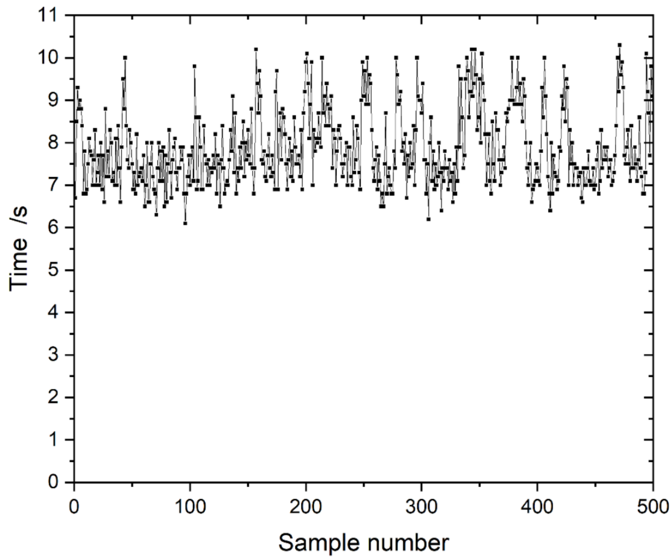

FIGURE 1. Distribution of 50-meter sprint scores among 500 students at a university in Nanjing

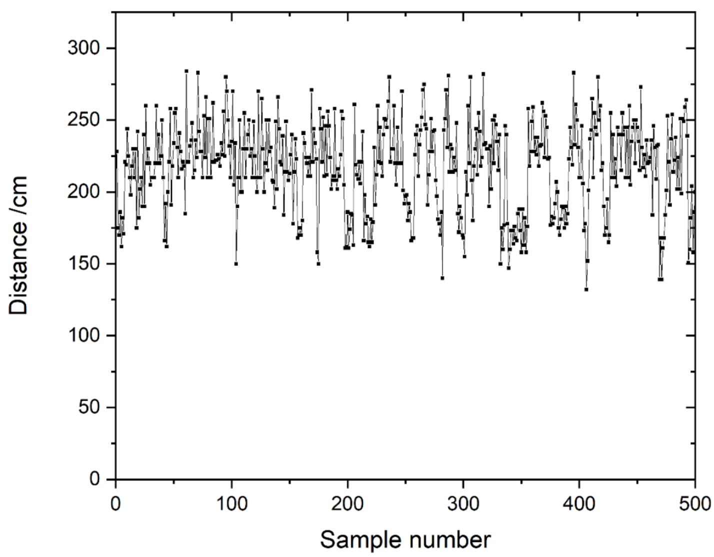

FIGURE 2. Distribution of standing long jump scores among 500 students at a university in Nanjing

is designed. The decision tree is employed as the first-stage algorithm, generating predicted physical test scores as its output. These predicted scores are then combined with the original student features (e.g., gender, height, weight, and vital capacity) to form an augmented input dataset for the second-stage multilayer perceptron (MLP). The MLP subsequently produces refined physical test score predictions. This hierarchical architecture effectively integrates the advantages of both algorithms: the decision tree's insensitivity to outliers and the MLP's capacity for nonlinear pattern recognition, thereby further enhancing the robustness of model predictions [13].

IV. RESULTS AND DISCUSSION

A. EVALUATION METRICS

Mean Squared Error (MSE): The average of squared differences between predicted values and true values [14,15].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

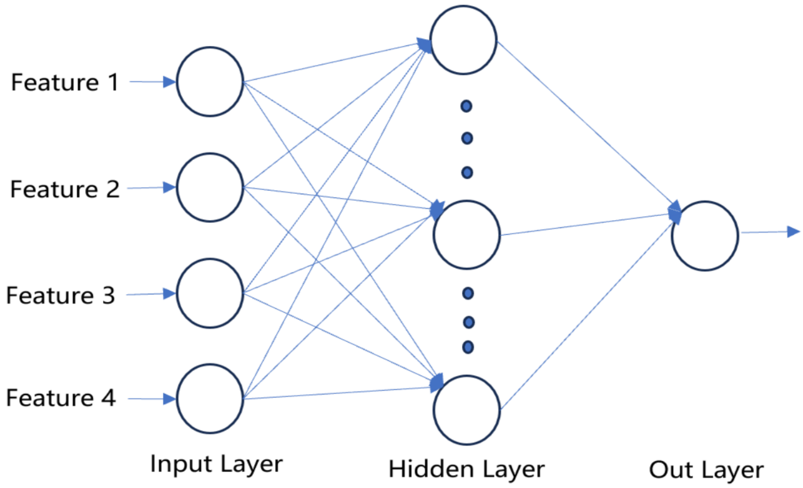

FIGURE 3. Schematic Diagram of the Multilayer Perceptron (MLP) Network Structure

Where $y_i, \hat{y}_i$ represent predicted value and true value separately.

Mean Absolute Error (MAE): The average of absolute differences between predicted values and true values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

Where $y_i, \hat{y}_i$ represent predicted value and true value separately.

B. MODEL TRAINING AND TESTING RESULTS

The study utilized physical fitness test data from 3,000 university students at a Nanjing-based institution. The dataset was split into training and test sets at a 4:1 ratio, with 20% of the training set allocated as a validation set. Experimental results demonstrated that the optimal Multilayer Perceptron (MLP) architecture comprised: an input layer (4 neurons receiving 4 features), two hidden layers (12 neurons each with ReLU activation), and an output layer (1 neuron). The Adam optimizer minimized the MSE loss function across 50 epochs with a batch size of 100. The training results are as follows: Table 1 compares the training, validation and test results of Multilayer Perceptron (MLP), Decision Tree, and Ensemble learning model on 50-meter sprint test scores. The Ensemble learning model demonstrated superior performance, achieving lower MSE and MAE values than both counterparts across all datasets.

TABLE 1. 50-meter sprint training, validation and testing results utilizing different learning models

Model	Training dataset		Validation dataset		Testing dataset	
	MSE	MAE	MSE	MAE	MSE	MAE
MLP	0.40	0.45	0.45	0.49	0.43	0.50
Decision tree	0.43	0.50	0.46	0.56	0.44	0.50
Ensemble learning model	0.26	0.4	0.30	0.42	0.30	0.41

Table 2 compares the training, validation and test results of Multilayer Perceptron (MLP), Decision Tree, and Ensemble learning models on standing long jump test scores. The Ensemble learning model demonstrated superior performance, achieving lower MSE and MAE values than both counterparts across all datasets.

TABLE 2. Standing long jump training, validation and testing results utilizing different learning models

Model	Training dataset		Validation dataset		Testing dataset	
	MSE	MAE	MSE	MAE	MSE	MAE
MLP	496	18	662	19	490	18
Decision tree	549	17	650	18	540	17
Ensemble learning model	334	13	460	15	340	13

C. FEATURE IMPORTANCE

In college student physical fitness tests, evaluating the feature importance of body measurement data (e.g., height, weight, and vital capacity) in machine learning models can reveal the relative contribution of these metrics to physical health assessment from a data-driven perspective. By ranking feature importance, we can quickly identify which metrics have the greatest impact on predicting overall physical fitness scores (or individual test items). This facilitates model optimization and provides practical guidance for fitness evaluation, intervention strategy formulation, and physical education improvement.

During model training, three anthropometric features including height, weight, and vital capacity of university students were utilized as input. Feature importance was quantitatively evaluated for both Multilayer Perceptron (MLP) and Decision tree models to assess their impact on model outputs, as illustrated in the Figure 4. Here, the abscissa represents physical characteristics such as height, weight, and vital capacity, while the ordinate indicates the normalized feature importance.

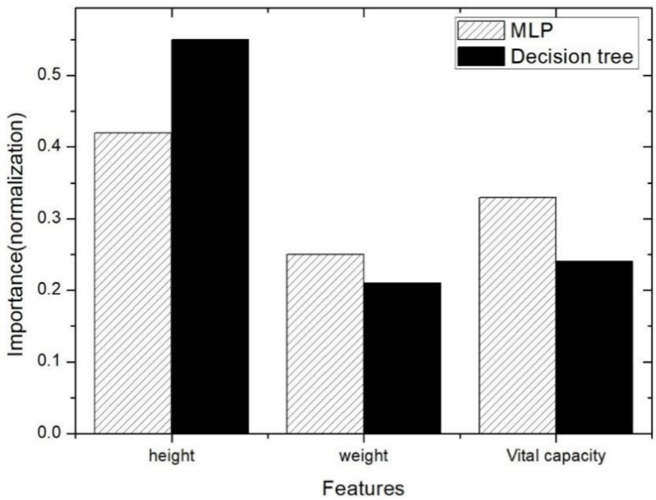

FIGURE 4. Distribution of feature importance in MLP and decision tree models

In Figure 4, the importance proporons of height, weight, and vital capacity in the MLP model are 0.42, 0.25, and 0.33 respectively while 0.55, 0.21, and 0.24 in the decision model. It is found that height is the most influential predictor in both models, with vital capacity ranking second in predictive significance. Based on the results, targeted interventions on students' weight and vital capacity can be effectively implemented to further enhance their physical fitness levels. During weight management interventions, BMI should serve as the primary metric, with comprehensive consideration given to both height and weight parameters.

V. APPLICATION OF RESULTS

A. ANOMALY ANALYSIS

Model-predicted anomalies are classified into two categories: correct predictions and incorrect predictions. For incorrect predictions, exclusion is required based on practical validation. Correct predictions hold practical value for physical fitness data analysis and should undergo the following protocol:

(1) Raw Data Validation: Verify measurement accuracy due to inevitable human errors in large-scale physical testing; Retest anomalies with validated data inconsistencies.

(2) Deviation Classification: Overestimated predictions: Indicate significant performance improvement potential, warranting personalized training programs; Underestimated predictions: Deemed non-instructive and excluded from further analysis.

Figure 5 presents a comparison between actual measured values (triangular markers) and predicted values (circular markers) from an ensemble learning model for 20 male 50-meter dash test results. The predicted values show good agreement with the actual measurements, with some outliers observed. Notably, data point no.13 within the dashed box exhibits a 10% underestimation (predicted: 7.2s vs actual: 7.92s). After excluding potential measurement errors, this outlier may indicate the student potential for performance improvement in the 50-meter dash. Therefore, targeted training interventions for the student could potentially yield significant performance enhancements.

B. ERROR CORRECTION OF FITNESS TEST DATA

Taking the 50-meter sprint scores of university students in Nanjing as an example, the ensemble learning model was applied to 4,000 samples, identifying 277 predictive outliers. After retesting these 277 students, 23 cases showed actual measurements that returned to normal predicted ranges, indicating prior measurement errors with an accuracy improvement of 0.58%. This method effectively enhances the measurement accuracy of fitness test data.

C. ANALYSIS OF PERSONALIZED TRAINING EFFECTS

The four-year physical fitness test scores of the 2020 freshman cohort from a university in Nanjing were selected as the research object. An ensemble learning model was employed to predict outliers and eliminate measurement errors. Students whose predicted scores were higher than their measured scores and who failed to reach the passing standard were assigned to the experimental group, while the remaining students with failing physical fitness test scores in the same grade were set as the control group.

As shown in Figure 6, taking male students as an example, 36 students who failed the 50-meter dash were tracked, and their average scores in this event over the four academic years (marked with circles in the figure) were used as the control group data. Their average scores from the first to the fourth year were 9.77, 9.76, 9.5 and 9.27 respectively. It can be observed that although their scores improved with the regular school training over the four years, the improvement effect was negligible, with a maximum increase of 5.1% from the lowest score of 9.77 to the highest of 9.27. The triangular points in the figure represent the

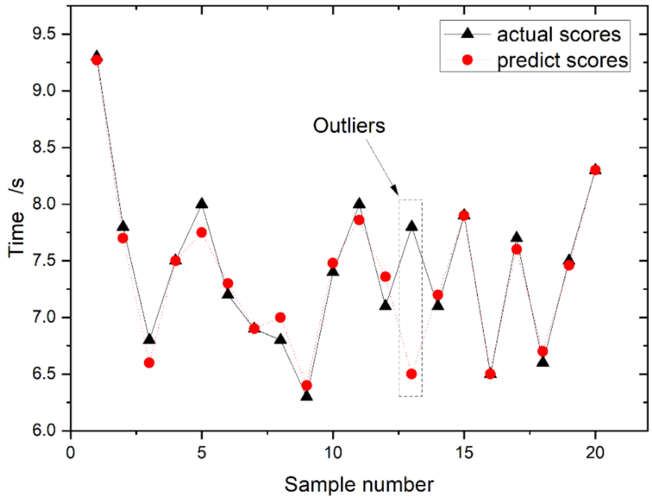

FIGURE 5. Predicted vs. actual 50-meter sprint scores: ensemble learning model performance

experimental group, which comprised 78 students identified as outliers from the 2020 physical fitness test data via the ensemble learning model. These students were tracked and received personalized training, with their 50-meter dash scores recording 9.75, 9.55, 8.8 and 8.9 from the first to the fourth year in sequence. The maximum improvement amplitude reached 9.74% from the lowest score of 9.75 to the best of 8.8, which was higher than that of the control group. Furthermore, all 78 male students in the experimental group met the passing standard (< 9.0) in their fourth year, whereas the maximum improvement in the average score of all male students in the same cohort during the same period was merely 0.53%.

Similarly, 93 female students who failed the 50-meter dash were tracked, and their average scores in this event across the four college years, marked with circles in the figure, were taken as the control group data. Their average scores from the first to the fourth year were 10.6, 10.35, 10.61 and 10.41 respectively. It can be observed that although their scores improved with regular school training over the four years, the improvement effect was negligible, with a maximum increase of merely 1.8% from the lowest score of 10.6 to the highest of 10.41. The triangular points in the figure represent the experimental group, which consisted of 48 female students identified as outliers from the 2020 physical fitness test data via the ensemble learning model. These students were tracked and received personalized training, with their scores changing to 10.55, 10.2, 9.5 and 9.7 in sequence from the first to the fourth year. The maximum improvement amplitude reached 9.95% from the lowest score of 10.55 to the best of 9.5, which was significantly higher than that of the control group. Furthermore, all 48 female students in the experimental group met the passing standard (< 10.2) in their fourth year, whereas the maximum improvement in the average score of all female students in the same cohort during the same period was only 1.1%.

VI. CONCLUSION

The implementation of the current physical fitness testing system for college students encounters challenges from multiple dimensions. First, the numerous testing items and large volume of data may make errors during execution. Issues such as insufficient standardization of testing procedures, inadequate personnel training, and variations in equipment precision collectively undermine data reliability, potentially causing results that fail to authentically reflect students' physical fitness levels. In this context, the integration of advanced textile-based wearable sensors represents a potential solution for improving the precision of data collection compared to traditional mechanical devices. Second, there are lack of effective intervention measures for students with substandard fitness. Within limited timeframes, personalized fitness improvement plans are lacking. Most institutions primarily focus on pass rates while neglecting individual differences, failing to develop targeted guidance strategies based on students' specific weaknesses in endurance, strength, or flexibility. This leads to limited effectiveness of corrective interventions.

In this paper, ensemble learning model is proposed which demonstrate significant guiding value for contemporary college students' physical fitness testing. Taking the 50-meter sprint as a case study, this research first employs historical data to train the weight parameters of the ensemble learning model. Subsequently, the trained model, combined with students' physical characteristics as input, is utilized to predict their 50-meter sprint performance.

A further analysis of prediction discrepancies reveals that experimental samples with notably higher predicted

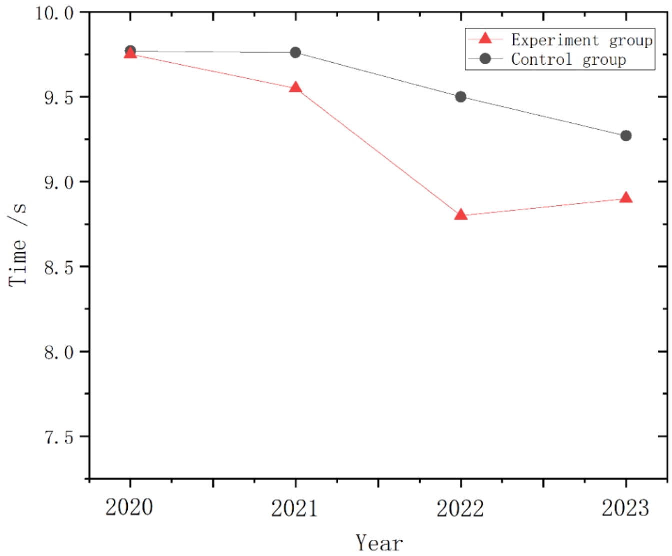

FIGURE 6. Comparison of 50-meter sprint performance between control and experimental groups in male undergraduates (class of 2020: freshman to senior year)

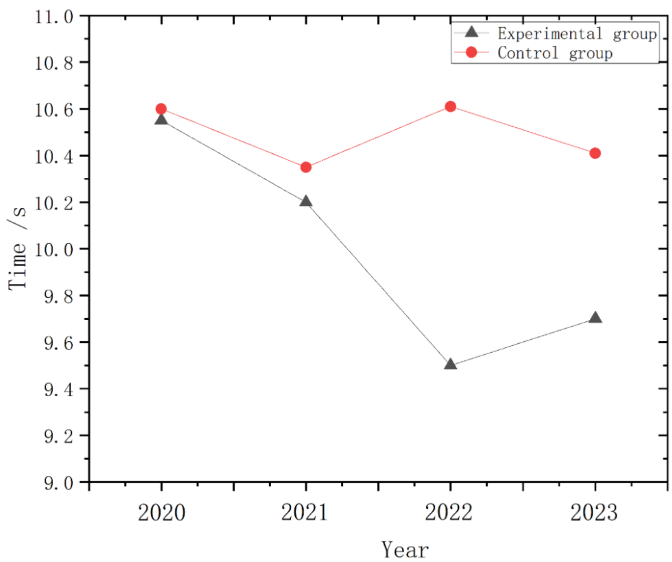

FIGURE 7. Comparison of 50-meter sprint performance between experimental and control groups in female undergraduates (class of 2020: freshman to senior year)

values were identified. These students received personalized training programs designed to fully exploit their potential. Tracking the physical fitness records of the 2020 cohort throughout their four-year college education, with all male and female students' performance data serving as reference samples, the experimental results demonstrate that both male and female experimental groups selected by the model achieved significantly greater improvements in 50-meter sprint performance compared to the control group after undergoing personalized training programs. This study provides valuable guidance for developing individualized physical training programs for college students, demonstrating considerable practical application value, particularly in providing data-driven support for the future design of functional athletic apparel tailored to individual performance needs. While this study only focused on the 50-meter sprint, future work should extend the framework to other test items and explore its integration with intelligent textile monitoring systems to further enhance the scientific management of student health.

AUTHOR CONTRIBUTIONS

writing-original draft preparation – Yanxia QIN; writing-review and editing – Mincong CHEN. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This work was supported by Special Project on Ideological and Political Work in Philosophical and Social Sciences Research of Colleges and Universities in Jiangsu, "Collaborative Research on Explicit and Implicit Ideological and Political Education in College Employment" (No.2022SJSZ0134). 2025 Annual University-Level Teaching Reform and Development Project of Nanjing Institute of Technology – "MY Camping" (JXIS2025MY42)

ACKNOWLEDGEMENTS

Not applicable.

HUMAN RESEARCH SUBJECTS

This study was approved by the Ethics Committee of Nanjing Institute of Technology, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

REFERENCES

[1] X. Lu, C. Yang, Y. Zhang, S. Huang, L. Li, H. Q. Chen, et al., “Test method for health-related physical fitness of college students in mobile internet environment,” Math Biosci Eng, vol. 16, no. 4, pp. 2189-2201, 2019, doi: 10.3934/mbe.2019107.

[2] Q. Zhang, W. Zhang, and F. Miao, “Intelligent recommendation method of an exercise program based on physical health data of college students,” Applied Artificial Intelligence, vol. 37, no. 1, pp. 1470-1488, 2023, doi: 10.1080/08839514.2023.2214458.

[3] H. Zhao, X. Yang, F. Qu, X. Zhang, L. Song, and X. Yang, “The impact of the COVID-19 outbreak on physical fitness and academic performance of Chinese college students,” J Am Coll Health, vol. 72, no. 5, pp. 1609-1616, 2024, doi: 10.1080/07448481.2022.2087472.

[4] D. A. Bonilla, I. A. Sánchez-Rojas, D. Mendoza-Romero, Y. Moreno, J. Kočí, L. M. Gómez-Miranda, et al., “Profiling Physical Fitness of Physical Education Majors Using Unsupervised Machine Learning,” Int J Environ Res Public Health, vol. 20, no. 1, pp. 146, 2022, doi: 10.3390/ijerph20010146.

[5] T. Cai, J. Long, J. Kuang, F. You, T. Zou, and L. Wu, “Applying machine learning methods to develop a successful aging maintenance prediction model based on physical fitness tests,” Geriatr Gerontol Int, vol. 20, no. 6, pp. 637-642, 2020, doi: 10.1111/ggi.13926.

[6] M. Rithani, R. P. Kumar, and S. Doss, “A review on big data based on deep neural network approaches,” Artificial Intelligence Review, vol. 56, no. 12, pp. 14765-14801, 2023, doi: 10.1007/s10462-023-10512-5.

[7] K. Xu and Z. Sun, “Predicting academic performance associated with physical fitness of primary school students using machine learning methods,” Complementary Therapies in Clinical Practice, vol. 51, Art. no. 101736, 2023, doi: 10.1016/j.ctcp.2023.101736.

[8] J. C. R. Whittington and R. Bogacz, “Theories of Error Back-Propagation in the Brain,” TRENDS IN COGNITIVE SCIENCES, vol. 23, no. 3, pp. 235-250, 2019, doi: 10.1016/j.tics.2018.12.005.

[9] J. R. QUINLAN, “Induction of decision trees,” Machine Learning, vol. 1, no. 1, pp. 81-106, 1986, doi: 10.1023/A:1022643204877.

[10] J. Wang, “The association between physical fitness and physical activity among Chinese college students,” Journal of American College Health, vol. 67, no. 6, pp. 602-609, 2019, doi: 10.1080/07448481.2018.1515747.

[11] B. Brown and C. Jones, “Physical fitness and mental health: A longitudinal study of college students,” International Journal of Mental Health Promotion, vol. 19, no. 2, pp. 58-67, 2021.

[12] M. M. Maldari, M. G. Jeanette, and J. R. David, “The impact of health education on physical activity correlates in college students,” Journal of American College Health, vol. 71, no. 1, pp. 111-116, 2023, doi: 10.1080/07448481.2021.1879812.

[13] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, “A survey on ensemble learning,” Frontiers of Computer Science, vol. 14, no. 2, pp. 241–258, 2019, doi: 10.1007/s11704-019-8208-z.

[14] T. Cao and Z. Liu, “Study on Automatic Recording and Health Promotion System of University Sports Results,” 2024 International Conference on Computers, Information Processing and Advanced Education (CIPAE); 26-28 Aug 2024; Ottawa, Canada. p. 843-847, doi: 10.1109/CIPAE64326.2024.00160.

[15] Y. Mou, L. Zhou, W. Chen, X. Zhao, Y. Liu, and C. Yang, “The analysis on college students’ physical fitness testing data — two cases study,” in Xu X, editor. 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 15-17 Dec 2017, Shenzhen, China. China: Harbin Institute of Technology (Shenzhen), 2017, pp. 252-255, doi: 10.1109/SPAC.2017.8304285.