A Hierarchical Classification Methodology for Online Key Courses on Ideological and Cultural Cultivation of Higher Education Students: A Perspective from Medical Colleges and Universities

Ziwei Zhou[1,#], Xuguang Dai[1,#], Biyu Huang[1], Manzhu Shen[1], Jiachen Liu[2]
[1]School of Medical Technology, Anhui Medical College, Hefei, Anhui, China
[2]School of Innovation and Entrepreneurship, Anhui Medical College, Hefei, Anhui, China
#Ziwei Zhou and Xuguang Dai contributed equally to this work and should be considered co-first authors

Abstract: In the ideological and cultural cultivation platform for vocational college students, there is a significant subjective bias in the classification of online key course resources, resulting in low accuracy of classification results. Therefore, this paper proposes a method for the classification of online key courses for the cultivation of higher vocational students' ideology and culture - from the perspective of medical colleges and universities. The Hadoop platform collects multidimensional data from the perspective of medical colleges and universities, combines web crawling technology, and stores it in HDFS after cleaning. Introducing Spark for real-time analysis and optimization of storage resource allocation. Build an online key course classification model and iteratively optimize the course classification results using clustering algorithms. Based on the classification model, the decision tree algorithm is used to classify the classified storage resources. By extracting and encoding features, constructing decision tree models, calculating information acquisition rates, and pruning, the classification of online key courses on ideological and cultural cultivation for vocational college students has been achieved. The experimental results show that this research method significantly improves the classification accuracy in the confusion matrix, optimizes the Kappa statistic, and thus improves the accuracy of the classification results.

Keywords: Ideological and Cultural Cultivation; Higher Vocational Students; Online Courses; Medical Colleges and Universities; Classification

1. Introduction

In the current context of the booming development of digital education, the widespread application of information technology has made the classification and mining of educational data increasingly important. These data are not only the key to improving teaching quality, but also an important means to optimize the allocation of educational resources. Especially in the field of higher education, facing massive and diverse educational data, how to accurately and efficiently achieve course-level classification has become an important issue in promoting the process of educational informatization. The rapid development of computer science, especially the rise of deep learning, machine learning and other technologies, has provided new ideas for the classification of educational data.

In existing research, Phuc & Chau [1] proposed an innovative classification method for heterogeneous educational data. By integrating data from different sources and formats, they achieved in-depth mining and accurate classification of course content. However, this method may encounter problems such as difficulty in feature selection and high computational complexity when dealing with highly heterogeneous data. Chen [2] used convolutional neural networks (CNNs) to classify large-scale data sets under the MapReduce framework, significantly improving classification efficiency. However, the complexity of the CNN model also brings about problems such as long training time and high resource consumption, and may not be accurate

 Copyright @ STEMM Institute Press

enough in capturing data features in specific fields. Kolchev et al [3] proposed a comprehensive method combining deep learning feature extraction and support vector machine (SVM) classification for the classification of breast ultrasound images. However, this method may still result in misclassification when dealing with edge cases or noisy data. Changguang Wang et al. [4] proposed a lightweight classification method for the detection and family classification of IoT malware. Although this method performs well in resource-constrained environments, the update speed and accuracy of its classification model may be affected in the face of increasingly complex malware variants.

In view of the shortcomings of existing research, this paper proposes a hierarchical classification method for online key courses in the ideological and cultural cultivation of higher vocational students. From the perspective of medical colleges and universities, considering the heterogeneity and complexity of data, combining the characteristics of the major and students, the algorithm and model are optimized in order to improve the classification accuracy while ensuring the classification efficiency.

2. Processing the Online Key Course Data of Ideological and Cultural Cultivation of Higher Vocational Students

The key online courses for the ideological and cultural cultivation of higher vocational students cover a variety of forms, including videos, documents, interactive discussions, etc. The data sources are wide and the formats are diverse. The data volume is huge, and there are problems such as redundancy and erroneous information [5-6]. In order to make it usable and valuable information, this paper first starts from the perspective of medical colleges and universities, recognizes the importance of these data for a comprehensive understanding of students' ideological dynamics and optimizing course cultivation programs, and then processes the course data.

From the perspective of medical colleges and universities, this paper selects Apache Hadoop as the core platform for processing the online key course data of the ideological and cultural cultivation of higher vocational students. Hadoop ensures the efficient management of massive course data with its excellent performance, high reliability and high scalability of open source distributed computing architecture [7-9]. Faced with the large number of students and complex data in the work of medical colleges and universities, Hadoop provides stable and efficient processing capabilities.

This paper uses web crawler technology to capture multi-dimensional data from the online course platform for the ideological and cultural cultivation of higher vocational students, including students' learning behavior, course content, and interaction data. These data can reflect students' learning participation, interests, and potential problems, which is crucial for optimizing course cultivation programs. By automating the data extraction process, this paper significantly improves processing efficiency and converts unstructured data into a structured format. The code example is shown in Figure 1.

```
from graphviz import Digraph
def create_data_processing_flowchart():
#Create a directed graph
Dot=Digraph (comment='Data Processing Process for Key Online Courses on Ideological and Cultural Cultivation of Vocational College Students')
#Add node
Dot.node ('A ',' Collect raw data ')
Dot.node ('B ',' Data Cleaning ')
Dot.node ('C ',' Store to HDFS')
Dot.node ('D ',' Apache Spark real-time analysis')
Dot.node ('E ',' Student Learning Behavior Pattern Recognition ')
Dot.node ('F ',' Course Resource Heat Analysis')

#Add Edge
dot.edges(['AB', 'BC'])
Dot. edge ('C ','D', label='Provide data through HDFS')
Dot. edge ('D ','E', label='scenarized application')
Dot. edge ('D ','F', label='scenarized application')
```

Figure 1. Code Examples for Capturing Data from Online Key Courses on Ideological and Cultural Cultivation for Higher Vocational Students

After cleaning, the raw data is stored in the Hadoop Distributed File System (HDFS). HDFS ensures data security through multi-node redundant storage, and the load balancing strategy achieves efficient data access [10-13]. Medical colleges and universities attaches great importance to data security because it involves students' personal information and learning conditions, and the consequences of leakage are serious. Efficient access ensures that staff can obtain data in a timely manner to support subsequent analysis and decision-making.

This paper also introduces Apache Spark as a data processing engine, and uses its memory computing capabilities to realize real-time analysis of course data. Real-time analysis can timely understand students' learning dynamics, such as course popularity and participation, and provide a basis for adjusting course content and teaching strategies. Through these technical means, this paper provides solid data support for the ideological and cultural cultivation of higher vocational students.

In view of the common quality defects in the original data, this paper constructs a data cleaning system specifically for online key courses on ideological and cultural cultivation of higher vocational students:

The missing value detection algorithm is used to identify data gaps, and statistical methods such as linear interpolation are used to complete the data. Assume that the data points of the online key course resources for the ideological and cultural cultivation of higher vocational students are (x_a, y_a), and its adjacent known data points are (x_b, y_b), the data points that need to be filled are x, the filling formula can be expressed by formula (1):

$$y = y_a + \frac{x \cdot y_b - x \cdot y_a - x_a \cdot y_b + x_a \cdot y_a}{x_b - x_a} \quad (1)$$

In formula (1), y indicates the completed key online course resources for the ideological and cultural cultivation of higher vocational students. For example, when processing the viewing time data of a teaching video in the online course of ideological and cultural cultivation of higher vocational students, if the viewing time data within a certain period of time is missing, this formula can be used to estimate the missing value based on the known viewing time data before and after. From the perspective of medical colleges and universities, complete data is the basis for accurate analysis, and the processing of missing values can ensure the integrity of the data and provide a reliable basis for subsequent decision-making.

This paper uses a hash algorithm to detect duplicate values, accurately identify and remove redundant records in the online course of ideological and cultural cultivation for higher vocational students, such as repeated assignments and comments in the discussion area, to ensure data quality and improve the authenticity of the analysis results.

In terms of data storage, this paper adopts a layered storage architecture to cache high-frequency access data to the Spark memory computing cluster to quickly respond to the work needs of medical colleges and universities; low-frequency historical data is archived to the HDFS cold storage layer to save storage resources. This architecture not only meets the use needs of data in different time dimensions, but also improves the efficiency of data storage and access. For example, the recent learning behavior data of the online course on ideological and cultural cultivation of higher vocational students will be frequently accessed, so it is cached to the Spark cluster to increase the access speed. The older historical data is archived to HDFS for emergency use.

3. Constructing a Grading Model for Key Online Courses

Based on the data of high-quality online courses on ideological and cultural cultivation of higher vocational students, this paper constructs a grading model, aiming to provide a scientific grading system for course resources, guide teaching practice, optimize resource allocation, and meet the work needs of medical colleges and universities.

The attribute vector of the key online courses C_i for the ideological and cultural cultivation of higher vocational students is $\overrightarrow{x_i} = (x_{i1}, x_{i2}, \ldots, x_{in})$, the attribute vector of course C_j is $\overrightarrow{x_j} = (x_{j1}, x_{j2}, \ldots, x_{jn})$, the weight vector is $\vec{w} = (w_1, w_2, \ldots, w_n)$, where $w_k \geq 0$ and $\sum_{k=1}^{n} w_k = 1$, then the calculation process for the weighted distance d_{ij} between the key online courses C_i for the ideological and cultural cultivation of higher vocational students and course C_j is shown in formula (2):

$$d_{ij} = \sqrt{\sum_{k=1}^{n} w_k \left(x_{ik} - x_{jk}\right)^2} \quad (2)$$

Where, x_{ik} and x_{jk} respectively represent the value of the key online courses C_i and course C_j of ideological and cultural cultivation for higher vocational students in attribute k, and w_k represents the weight of attribute k.

 Copyright @ STEMM Institute Press

The online key course data set of high-quality higher vocational students' ideological and cultural cultivation is $D=\{C_1,C_2,\ldots,C_m\}$, and the clustering result is $C=\{C_{11},C_{12},\ldots,C_{1k1};C_{21},C_{22},\ldots,C_{2k2};\ldots;C_{T1},C_{T2},\ldots,C_{TkT}\}$, namely the online key course data set D for the ideological and cultural cultivation of higher vocational students is divided into T cluster, the t th clusters contain k_t courses. Clustering objective function J is defined as formula (3):

$$J=\sum_{t=1}^{T}\sum_{C_i\in C_t} d_{ij} \tag{3}$$

By minimizing J, it can make the online key courses for ideological and cultural cultivation of higher vocational students in the same group under the same grade higher similarity, and make the courses between different grade groups more different.

In this paper, the data set contains key online courses on the cultivation of higher vocational students' ideology and culture, and each object has multi-dimensional attributes. By calculating the center of each type of data (i.e., the feature mean), and using formula (2) to calculate the Euclidean distance between the cluster center and the data point, the data point is assigned to the nearest cluster. Through loop iteration, the cluster to which the course belongs is continuously adjusted to optimize the classification results. When the change amplitude of clustering objective function J is less than the preset threshold, the algorithm terminates and outputs the final clustering result, thus realizing the grading of online key courses for the ideological and cultural cultivation of higher vocational students.

4. Implement Online Key Course Classification

From the perspective of medical colleges and universities, we can achieve course classification. The goal is to subdivide the graded courses according to business logic and actual needs, and provide targeted learning resources and teaching guidance for students of different levels and needs. For example, we can provide courses with different levels of difficulty and content for students with different professional backgrounds and learning foundations to meet the diverse needs of students and improve their learning effects and literacy. At the same time, classification helps medical colleges and universities understand the distribution of course resources, optimize resource allocation, and improve utilization efficiency.

The decision tree algorithm is applied to classify the key online courses for the ideological and cultural cultivation of higher vocational students. The decision tree consists of nodes and branches. The nodes test the characteristic attributes of the course data, the branches represent the test results, and the leaf nodes store the classification results [14-16]. Through layer-by-layer testing, the classification results of the courses are finally obtained by reaching the leaf nodes.

The implementation steps are as follows:

The classified online key course data of higher vocational students' ideological and cultural cultivation is used as the input data of the classification algorithm to extract and encode the data. For example, the classification rules such as professional relevance, learning difficulty, and course format are converted into numerical features so that the algorithm can process them.

Use the training data set to build a decision tree model. During the construction process, select appropriate features as partition nodes, such as the professional relevance of the course, learning difficulty, etc., and divide the data set into different subsets according to different feature values until the stopping condition is met (such as the data in the subset belongs to the same category or the maximum tree depth is reached). In the actual application of medical colleges and universities, the features and partitioning methods can be reasonably selected according to the importance of the classification rules and the data distribution to improve the accuracy of classification. For example, for students majoring in medicine, courses related to medicine can be given priority for partitioning.

Information gain rate is an important indicator used in decision tree algorithm to select the optimal partitioning features. In the classification of online key courses for ideological and cultural cultivation of higher vocational students, the information gain rate can be expressed by formula (4):

$$L(H)=\frac{Z(H)}{G(H)} \tag{4}$$

In formula (4), $Z(H)$ represents the information gain in the classification of online key courses for the ideological and cultural cultivation of higher vocational students for characteristics H, $G(H)$ represents the inherent value in the classification of online key courses for ideological and cultural cultivation of higher vocational students for characteristics H. The higher the information gain rate, the greater the contribution of the feature to the classification [17-18]. By calculating the information gain rate, this paper selects the optimal features to construct decision tree nodes, ensuring a small number of nodes and high recognition accuracy. When the decision tree is generated, there may be noise and outliers in the training set, which will affect the prediction accuracy. Therefore, this paper adopts the post-pruning method to cut off the redundant branches and retain the leaf nodes to form a new decision tree. The mining process is shown in Figure 2. The new tree is used to judge the online key course examples of ideological and cultural cultivation of higher vocational students.

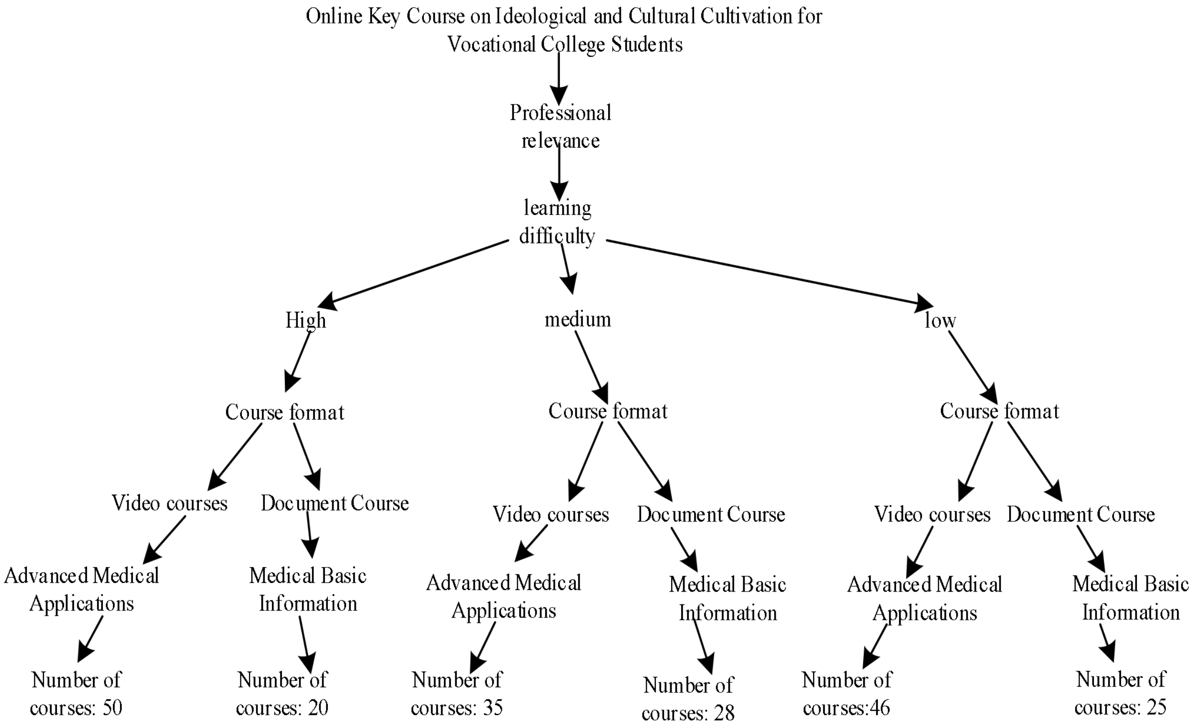

Figure 2. Schematic Diagram of the Classification and Mining Process of Online key Courses for the Ideological and Cultural Cultivation of Higher Vocational Students Based on the Decision Tree Algorithm

This paper uses the decision tree algorithm to test the information gain feature attributes of online key courses for the ideological and cultural cultivation of higher vocational students, selects the maximum gain to build nodes, and ensures that the number of nodes is small and the recognition is accurate. Branches are constructed according to the feature values, and instances are processed into subsets of the same type. Through pruning, the generalization ability of the decision tree model in the classification of online key courses for the ideological and cultural cultivation of higher vocational students is further improved, so that it can achieve better classification results when processing new course data [19-20] . Therefore, this paper completes the hierarchical classification method of online key courses for the ideological and cultural cultivation of higher vocational students - a study from the perspective of the work of medical colleges and universities.

5. Experiments

This paper carefully prepares the hardware and software environment for the grading and classification experiment of online key courses for the ideological and cultural cultivation of higher vocational students. The configuration is shown in Table 1.

The experimental data set is shown in Table 2:

This paper uses confusion matrix and Kappa statistic as core indicators. The confusion matrix

 Copyright @ STEMM Institute Press

intuitively displays the comparison between classification results and true labels, reflecting key performance such as recognition accuracy and misjudgment rate. Kappa statistic quantifies model consistency, eliminates the impact of uneven category distribution, and more objectively reflects model stability and reliability.

The hierarchical classification method proposed in this paper is set as the experimental group, Heterogeneous Educational Data Classification at the Course Level [1], Deep-learning feature extraction with their subsequent selection and support vector machine classification of the breast ultrasound images [2], A Lightweight IoT Malware Detection and Family Classification Method [3] As comparison method 1, comparison method 2 and comparison method 3 respectively.

Table 1. Experimental environment hardware and software configuration table

Serial Number	Category	Specific parameters
(1)	Operating system	Ubuntu 20.04 LTS
(2)	Data analysis tools	Python 3.11 (including Pandas and Scikit learn libraries)
(3)	The server	CPU: Intel Xeon Gold 6338 (2.0GHz, 32 cores)
		Memory: 256GB DDR4 ECC REG
		Storage: 8TB NVMe SSD (RAID 5)
(4)	Teacher Workstation	CPU: Intel Core i9-13900K (24 cores, 32 threads)
		Graphics card: NVIDIA RTX 4090 24GB
		monitor: 4K@144Hz (Dual screen)
(5)	Student Terminal	CPU: Intel Core i7-13700H (14 cores and 20 threads)
		Memory: 32GB DDR5
		Graphics card: NVIDIA RTX 3060 6GB
(6)	network device	Core Switch: Huawei V735-S48T4X (48 Gigabit+4 Gigabit SFP+)
		Wireless AP: Huawei AP8082DN-E8 (supports Wi Fi 6E)

Table 2 Experimental dataset information

Serial Number	Dataset category	Detailed description
(1)	Student Basic Information	Contains basic information such as student name, student ID, gender, age, major, etc., for student identification and classification analysis.
(2)	Course learning data	Record students' learning behavior data in various online courses, such as course viewing time, participation in discussions, homework completion status, etc.
(3)	Course grading data	Classify online courses based on factors such as difficulty and depth of content.
(4)	Student feedback data	Collect students' feedback on online courses, including course satisfaction, learning experience, etc.
(5)	Teacher evaluation data	Teachers evaluate students' performance in online courses, including learning attitude, participation, homework quality, etc.

Four methods were used to classify and process the online key courses of ideological and cultural cultivation for higher vocational students. After the experiment, the confusion matrix results corresponding to the four methods were obtained, as shown in Figure 3.

As can be seen from Figure 3, the proposed method has obvious advantages. In terms of classification accuracy, the proposed method predicts the number of correct samples in each category far exceeding other comparative methods. Especially in the core category classification, the main diagonal elements of the confusion matrix are prominent, indicating that it can accurately capture the core characteristics of the online course of ideological and cultural cultivation of higher vocational students and achieve accurate classification. At the same time, the proposed method has a low misclassification rate and small values of non-diagonal elements, which effectively avoids confusion in course categories. This is due to the fact that the proposed method incorporates a comprehensive and detailed feature extraction and classification strategy, fully considering the complexity and diversity of the course, and ensuring better classification results.

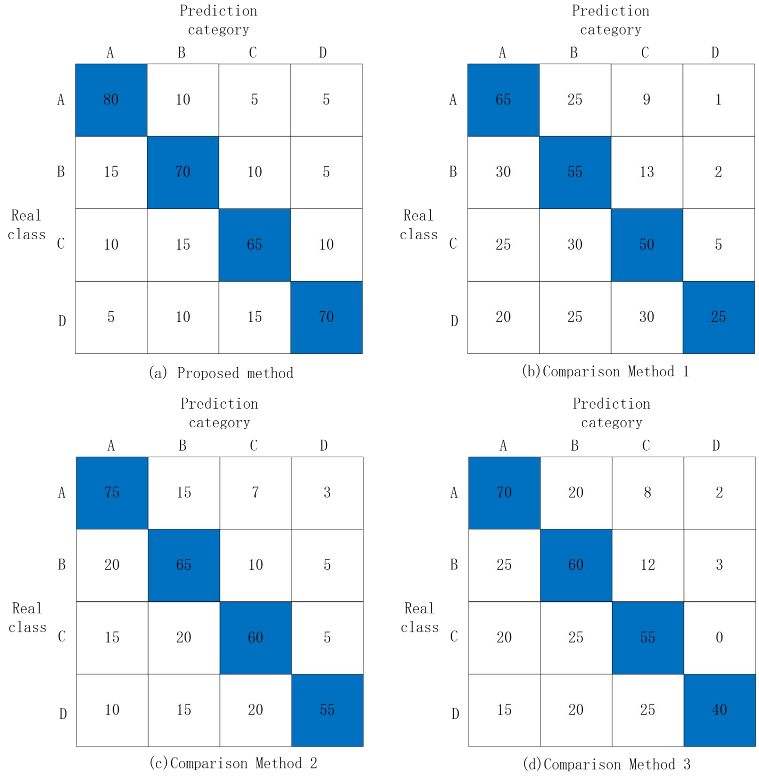

Figure 3. Confusion Matrix Comparison Results of Four Methods

This paper further uses the Kappa statistic to quantitatively compare the four methods, and the comparison results are shown in Figure 4.

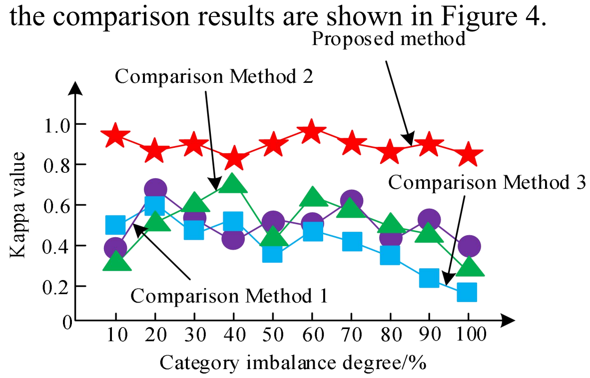

Figure 4. Kappa Statistic Comparison Results of Four Methods

As can be seen from Figure 4, the proposed method has a significant advantage in Kappa statistics. When dealing with the grading and classification of online courses for the cultivation of higher vocational students' ideology and culture, the proposed method performs better. This advantage is mainly due to the fact that the proposed method fully considers the diversity and hierarchy of course content. By constructing a decision tree and performing pruning operations, it can continuously classify the graded online key courses for the cultivation of higher vocational students' ideology and culture, thereby improving the accuracy of the classification results.

6. Conclusion

This paper uses Hadoop and Spark technology to automatically capture, clean, convert and efficiently store massive multi-source heterogeneous course data, ensuring the integrity and security of the data. A grading model based on weighted distance and clustering objective function is constructed, and iterative optimization is used to achieve the scientific grading of online key course resources for the ideological and cultural cultivation of higher vocational students. The decision tree algorithm is combined to finely classify the grading courses. Through feature extraction, model construction and other steps, the classification accuracy and generalization ability are significantly improved, providing strong support for teaching practice. In future research, federated learning technology can be introduced to allow medical college student union committees to share dynamic data of medical students' thoughts while protecting data privacy. A distributed dynamic classification model can be constructed to achieve adaptive updates of classification rules as the learning situation evolves.

Acknowledgements

This work was supported by the Anhui

 Copyright @ STEMM Institute Press

Provincial Higher Education Research Project (No. 2022AH052312, No.2023AH052579); Annual Ideological and Political Theory Course Teaching and Research Project of Anhui Medical College(No.2023ahyzsxzzjy01); Research Foundation for Advanced Talents of Anhui Medical College(No.2023RC007).

References

[1] Nguyen Hua Gia Phuc, Vo Thi Ngoc Chau. Heterogeneous Educational Data Classification at the Course Level. VIETNAM JOURNAL OF COMPUTER SCIENCE, 2021, 8(3):337-355.
[2] Min Chen. Classification with Convolutional Neural Networks in MapReduce. JOURNAL OF COMPUTER AND COMMUNICATIONS, 2024, 12(8): 174-190.
[3] Kolchev A A, Pasynkov D V, Egoshin I A, et al. Deep-learning feature extraction with their subsequent selection and support vector machine classification of the breast ultrasound images. Компьютерная оптика, 2024, 48(5): 753-761.
[4] Changguang Wang, Ziqi Ma, Qingru Li, et al. A Lightweight IoT Malware Detection and Family Classification Method. JOURNAL OF COMPUTER AND COMMUNICATIONS, 2024, 12(4): 201-227.
[5] Lamba, Monika,Gigras, et al.Classification of plant diseases using machine and deep learning.OPEN COMPUTER SCIENCE, 2021, 11(1): 491-508.
[6] Chen X, Zou D, Xie H, et al. Automatic Classification of Online Learner Reviews Via Fine-Tuned BERTs. International Review of Research in Open and Distributed Learning, 2025, 26(1): 57-79.
[7] Deshpande, Nilkanth Mukund, Gite, et al. A New Method for Diagnosis of Leukemia Utilizing a Hybrid DL-ML Approach for Binary and Multi-Class Classification on a Limited-Sized Database. CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES, 2024, 139(1): 593-631.
[8] Ba S, Hu X, Stein D, et al. Assessing cognitive presence in online inquiry-based discussion through text classification and epistemic network analysis. British Journal of Educational Technology, 2023, 54(1): 247-266.
[9] Attiapo Acybah Morel Omer.Image Classification Based on Vision Transformer. JOURNAL OF COMPUTER AND COMMUNICATIONS, 2024, 12(4): 49-59.
[10]Pathak D, Kashyap R. Neural Correlate-Based E-Learning Validation and Classification Using Convolutional and Long Short-Term Memory Networks. Traitement du Signal, 2023, 40(4): 1457-1467.
[11]Xiaonan Li,Qiang Wang,Conglai Fan,et al. A New Malicious Code Classification Method for the Security of Financial Software. COMPUTER SYSTEMS SCIENCE & ENGINEERING, 2024, 48(3): 773-792.
[12]Nayak P, Vaheed S, Gupta S, et al. Predicting students' academic performance by mining the educational data through machine learning-based classification model. Education and Information Technologies, 2023, 28(11): 14611-14637.
[13]Li, Ming, Fang, Zhong, Cao, Wanwan, et al. Residential Electricity Classification Method Based On Cloud Computing Platform and Random Forest. COMPUTER SYSTEMS SCIENCE AND ENGINEERING, 2021, 38(1): 39-46.
[14]Zhong X. Text classification and topic mining of intelligent forum comments in university MOOC based on CNN networks. International Journal of Computational Systems Engineering, 2025, 9(1): 58-67.
[15]Luo X, Zhao Y, Qin Y, et al. Towards semi-supervised universal graph classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 36(1): 416-428.
[16]Zarina T. LEXICAL-SEMANTIC CLASSIFICATION OF ONLINE SHOPPING TERMS. AMERICAN JOURNAL OF EDUCATION AND LEARNING, 2025, 3(4): 433-438.
[17]Baruth O, Cohen A. Personality and satisfaction with online courses: The relation between the Big Five personality traits and satisfaction with online learning activities. Education and Information Technologies, 2023, 28(1): 879-904.
[18]Liu Z, Kong X, Chen H, et al. MOOC-BERT: Automatically identifying learner cognitive presence from MOOC discussion data. IEEE Transactions on Learning Technologies, 2023, 16(4):

528-542.
[19]Liu Z, Kong X, Liu S, et al. Effects of computer-based mind mapping on students' reflection, cognitive presence, and learning outcomes in an online course. Distance Education, 2023, 44(3): 544-562.
[20]Jiang Y, Gan L. Research on College Physical Education Teaching Practice Based on Classification and Stratification of Health Fitness. Journal of Asian Society for Health & Exercise, 2023, 5(1): 1-14.

 Copyright @ STEMM Institute Press