Applied & Educational Psychology (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/appep.2026.070107
ISSN 2523-5842 Vol. 7 Num. 1

Integrating Sandplay Analysis and CNN–SVM Modeling for Predicting English Learning Difficulties among Vocational College Students: A Computational Psychology Approach

Lingjie Wu[1,a], Zhenyi Liu[1,b], Duoduo Yu[2,c,*], Xi Wang[3,d], Shaochen Lin[4,e], Jianhan Peng[5,f], Dugui Chen[6,g]

[1]School of Health Sciences, Wenshan Vocational and Technical College, Wenshan, China
[2]School of General Education, Wenshan Vocational and Technical College, Wenshan, China
[3]School of Medicine and Nursing, Shandong Vocational University of Foreign Affairs, Weihai, China
[4]Department of Geriatrics, Yunnan Provincial Psychiatric Hospital, Kunming, China
[5]School of Teacher Education, Wenshan University, Wenshan, China
[6]School of Artificial Intelligence, Wenshan Vocational and Technical College, Wenshan, China
[a]45692517@qq.com, [b]liuzhenyi1997@163.com, [c]1299052890@qq.com, [d]88654297@qq.com, [e]658342972@qq.com, [f]74562987@qq.com, [g]35246743@qq.com
**Corresponding author*

Keywords: Sandplay analysis; learning difficulties; convolutional neural network; support vector machine; computational psychology

Abstract: This study integrated sandplay analysis with a hybrid convolutional neural network–support vector machine (CNN–SVM) framework to identify English learning difficulties among vocational college students, with the aim of improving the objectivity and reproducibility of projective assessment in educational psychology. A total of 190 students were included, of whom 134 were classified as having English learning difficulties and 56 as non-difficulty cases. Group assignment followed a dual-criterion approach: students with semester English scores below 60, low academic self-efficacy, and high foreign language anxiety were assigned to the difficulty group, whereas those with scores above 70, high self-efficacy, and low anxiety were assigned to the non-difficulty group. Under standardized and controlled conditions, all participants completed a sandplay task, and the resulting images were preprocessed through cropping, normalization, and resizing, with data augmentation applied to mitigate class imbalance. A shallow CNN was used to extract symbolic and spatial features from the sandplay images, and the resulting feature embeddings were then classified using an SVM with a radial basis function kernel. Model hyperparameters were optimized through nested cross-validation, and robustness was further evaluated using repeated hold-out validation. The CNN–SVM model achieved a mean validation accuracy of 91.5% and an area under the curve (AUC) of 0.95. On the independent test set, the model reached an accuracy of 83.9%, and recall for the difficulty group was 96.2%, indicating strong sensitivity and a reduced likelihood of missing at-risk students. Grad-CAM visualizations further showed that figurine clustering and central spatial arrangements were among the most important discriminative cues, consistent with established interpretations of sandplay symbolism. Overall, the findings suggest that

quantifiable symbolic cues in sandplay can support accurate prediction of learning difficulties, and that this reproducible framework offers a promising bridge between projective techniques and artificial intelligence for early screening. Future research should include larger and more diverse samples, as well as broader assessment indicators, to strengthen the external validity of the model.

1. Introduction

Learning difficulties, particularly in second language acquisition, represent a persistent challenge among vocational college students. In China, English is a compulsory subject, and proficiency in English is often linked not only to academic success but also to broader career opportunities [1]. Students with English learning difficulties frequently experience lower academic self-efficacy, heightened language anxiety, and reduced motivation, which in turn may exacerbate their risk of academic underachievement and psychological distress [3].

Traditional approaches to identifying learning difficulties have relied heavily on academic performance scores and psychometric instruments such as self-report questionnaires and teacher assessments [9]. While these methods provide valuable insights, they are subject to self-report biases, cultural influences, and inter-rater variability, which limit their objectivity and replicability[2]. This underscores the urgent need for more reliable, automated, and scalable approaches to identifying students at risk.

Projective techniques, including drawing assessments and sandplay, have been widely applied in counseling and educational psychology to explore individuals' implicit psychological states[8].Among these methods, sandplay is particularly valued for its capacity to externalize inner experiences through symbolic representation, spatial configuration, and the use of figurines[10]. Nevertheless, conventional interpretations of sandplay remain predominantly qualitative and heavily reliant on expert judgment, which raises concerns regarding subjectivity and limited reproducibility[4].

Recent developments in artificial intelligence (AI) and machine learning provide new opportunities to overcome these challenges. Convolutional neural networks (CNNs) are well recognized for their ability to extract hierarchical features from images, whereas support vector machines (SVMs) demonstrate strong performance in high-dimensional tasks with relatively small datasets[5]. Combining these approaches into a CNN–SVM framework offers the potential to transform symbolic projective methods into quantifiable and replicable computational analyses. Applied to sandplay images, this hybrid strategy enables the identification of latent features associated with learning difficulties in a manner that is both objective and reproducible[6].

The present study therefore integrates sandplay analysis with a CNN–SVM hybrid model to predict English learning difficulties among vocational college students[7]. This work seeks to advance methodological innovation in computational psychology while providing a practical screening tool for early detection of at-risk learners

2. Research Methods

The data used in this study were obtained from standardized sandplay sessions conducted with vocational college students, rather than from raw, uncontrolled collections. Specifically, the data acquisition process was carried out between September 2024 and January 2025, and the final

dataset was officially incorporated into model construction on March 1, 2025. A total of 190 students participated in the sandplay sessions.Based on a dual-criterion approach, students were classified into groups using both academic performance and psychological assessments. Specifically, those with English semester grades below 60 out of 100, combined with low self-efficacy and high foreign language anxiety, were included in the English-learning difficulty group, whereas students with grades above 70, high self-efficacy, and low anxiety were assigned to the non-difficulty group. This procedure ensured that the labeled supervised dataset reflected not only objective achievement but also relevant psychological dimensions.

Each student completed a sandplay task under standardized conditions with identical sand trays and figurines. Images of the sand trays were captured using fixed lighting and angle settings to ensure consistency. All images were preprocessed through cropping, resizing, and normalization to meet the input requirements of the model.

A convolutional neural network (CNN) was employed for feature extraction, where the pixel matrix of the sandplay image served as the independent variable (X), and the categorical group label (difficulty vs. non-difficulty) as the dependent variable (y). The deep feature representations extracted by the CNN were subsequently used as inputs for a support vector machine (SVM) classifier, which employed a radial basis function (RBF) kernel to optimize the classification boundaries.

The constructed CNN-SVM model was evaluated using standard machine learning criteria. Performance metrics included confusion matrices, receiver operating characteristic (ROC) curves, and the area under the ROC curve (AUC). To ensure robustness, cross-validation was conducted to assess the model's stability across different data splits.

2.1. Data Collection

This study involved a total of 190 vocational college students. Participant grouping was established through a dual-criterion approach that combined academic achievement with psychological assessment. In the first step, semester English course scores were used: students scoring below 60 out of 100 were placed in the difficulty group, while those scoring above 70 were assigned to the non-difficulty group. To reduce heterogeneity and sharpen the contrast between groups, students whose scores fell within the 60–70 range were excluded. In the second step, two widely used psychological instruments were administered: the General Self-Efficacy Scale (GSES) or Academic Self-Efficacy Scale, which measure learners' confidence in their academic abilities, and the Foreign Language Classroom Anxiety Scale (FLCAS), which assesses anxiety specifically in English language learning. Students were categorized as having English learning difficulties only if they met both criteria—low academic scores (<60) in combination with low self-efficacy and high anxiety. In contrast, students with high academic scores (>70) together with high self-efficacy and low anxiety were classified as non-difficulty cases.

In educational psychology, the extreme-group design is frequently adopted to highlight contrasts by selecting participants from the upper and lower ends of the distribution[12]. Such an approach reduces group overlap, increases statistical power, and strengthens construct validity by considering not only objective performance outcomes but also psychological characteristics. In the Chinese education system, a score of 60 is officially regarded as the passing threshold [13]. Students who consistently perform below this benchmark are typically classified as experiencing "learning difficulties" in domestic research[14]-[15]. By supplementing academic criteria with validated psychological measures, the present study provides a classification framework that reflects the multifaceted nature of learning difficulties, consistent with current perspectives in educational and developmental psychology[17].

Given the unequal group sizes (134 vs. 56), the dataset presented a class imbalance problem. To mitigate its impact on model training and evaluation, oversampling techniques were applied to the minority class (non-difficulty group), including image rotation and flipping. This procedure increased the representation of non-difficulty samples and reduced bias in classification. Additionally, evaluation metrics such as precision, recall, F1-score, and AUC were reported using macro-averaging, ensuring that performance assessment was not dominated by the majority class.

All participants completed a standardized sandplay task in a controlled environment. Identical sand trays and figurines were provided, and each student was asked to freely construct a scene within the sand tray. The sessions were conducted under uniform conditions of lighting and camera angle to ensure consistency across participants. Images of the completed sand trays were then collected and archived to form the experimental dataset.

Although the study was limited to vocational college students within a specific age range (16–17years), this population was intentionally selected. In China, students in vocational colleges often face substantial academic challenges, lower self-efficacy in foreign language learning, and heightened psychosocial pressures related to future employment. Previous national reports have highlighted that English learning difficulties are closely associated with anxiety, avoidance, and negative self-concept among this demographic. Focusing on this group therefore addresses a pressing educational and psychological issue within a contextually significant population.

2.1.1. Inclusion Criteria

Eligible participants were vocational college students aged 16–17 years, both male and female, who had completed at least one semester of English coursework. Grouping was determined through a dual-criterion approach that combined academic performance with psychological assessment.

Students were classified as belonging to the English learning difficulty group if they obtained an average English score below 60/100 and simultaneously demonstrated low self-efficacy (measured by the General Self-Efficacy Scale or Academic Self-Efficacy Scale) together with high language anxiety (measured by the Foreign Language Classroom Anxiety Scale, FLCAS).

Conversely, students were assigned to the non-difficulty group if their average English score was above 70/100, accompanied by high self-efficacy and low language anxiety.

Those with scores between 60 and 70, or whose psychological scale results did not meet the defined criteria, were excluded to ensure a clear separation between groups. Only students who gave informed consent and voluntarily took part in the sandplay sessions conducted under standardized conditions were included.

2.1.2. Exclusion Criteria

Students were excluded if they had serious physical conditions (e.g., cardiovascular, hepatic, or renal diseases) that could affect participation, or if they had a previously diagnosed psychiatric disorder not directly related to learning difficulties (such as schizophrenia or severe generalized anxiety). Individuals who had undergone prolonged psychotherapy or pharmacological treatment before the study were also excluded. In addition, students unwilling to participate or unable to provide informed consent were not eligible.

The age range (16–17years) was deliberately chosen to minimize variability due to developmental differences and to focus on the specific challenges faced by vocational college students. Gender distribution was relatively balanced across the sample, ensuring representativeness. The basic demographic characteristics of the participants are summarized in Table 1.

A total of 134 vocational college students with English learning difficulties were included in the experimental group. Group assignment was determined using both academic and psychological

criteria: students with English scores below 60, low self-efficacy, and high language anxiety were classified as having learning difficulties. A total of 56 students with scores above 70, high self-efficacy, and low anxiety were assigned to the non-difficulty group. Students scoring between 60 and 70 or not meeting the psychological criteria were excluded.It should be noted that students with English scores falling between 60 and 70, or those not meeting the psychological scale criteria, were excluded to guarantee a clear separation between groups and to enhance construct validity in Table 2.

Table 1. Sex distribution of vocational college students

Sex	Difficulty group (n=134)	Non-difficulty group (n=56)	Total (n=190)	Percentage (%)
Male	72	30	102	53.7
Female	62	26	88	46.3
Total	134	56	190	100.0

Table 2. Age distribution of vocational college students

Age	Difficulty group	Non-difficulty group	Total	Percentage (%)
16	64	28	92	48.40
17	70	28	98	51.6
Total	134	56	190	100.0

2.2. Study Hypotheses

H1: Feature representations extracted through a convolutional neural network (CNN), when combined with support vector machine (SVM) classification, will effectively differentiate vocational college students with English learning difficulties from their peers without such difficulties, using sandplay images as input.

H2: Distinct visual characteristics within sandplay constructions—such as the choice of figurines, spatial organization, and symbolic configurations—are expected to correlate with the severity of English learning difficulties and may serve as valid indicators of underlying psychological states.

H3: Grad-CAM visualizations can identify key regions and symbolic elements within sandplay images that contribute most strongly to the classification process, thereby providing interpretable links between the model's decisions and underlying psychological patterns.

2.3. Study Materials

To ensure that group assignment reflected not only academic achievement but also relevant psychological dimensions, two standardized self-report measures were administered alongside English course grades.

2.3.1. Academic Self-Efficacy

Students' confidence in their learning ability was evaluated using either the General Self-Efficacy Scale[16] or the Academic Self-Efficacy Scale, depending on institutional use. The GSES consists of 10 items rated on a 4-point Likert scale (1 = "not at all true" to 4 = "exactly true"). Higher scores indicate stronger beliefs in one's capacity to manage academic challenges effectively. Previous studies have shown that the scale demonstrates solid psychometric properties among Chinese adolescents, with Cronbach's α values typically reported above 0.85 (Wang, C. K).Both the Chinese versions of the General Self-Efficacy Scale and the Foreign Language Classroom Anxiety Scale[18]were used in this study. The English translations of these instruments are provided in

Supplementary Materials 1–2 for reference.

2.3.2. Foreign Language Classroom Anxiety

English learning anxiety was assessed using the Foreign Language Classroom Anxiety Scale (FLCAS), originally developed by Horwitz and colleagues. This instrument contains 33 items scored on a 5-point Likert scale (1 = “strongly disagree” to 5 = “strongly agree”). It measures three key aspects of foreign language anxiety: communication apprehension, test anxiety, and fear of negative evaluation. Higher overall scores represent greater language-related anxiety. The FLCAS has been widely applied in research on Chinese English learners, and its internal consistency is well established, with Cronbach’s α values often exceeding 0.90 [11].

By integrating objective academic grades with validated psychological scales, the present study adopted a multidimensional framework for defining English learning difficulties, consistent with contemporary perspectives in educational and developmental psychology.

The study materials consisted of standardized sandplay sets, digital imaging devices, and computational resources for model training.

Sandplay sets: Each participant was provided with a uniform sand tray (57 × 72 × 7 cm) filled with fine white sand. A standardized collection of figurines and miniature objects (e.g., animals, buildings, vehicles, natural elements, and human figures) was available for all students. These items were selected to ensure a wide range of symbolic representations while maintaining consistency across participants.

Image acquisition: Upon completion of each sandplay construction, high-resolution digital photographs were taken under controlled conditions of fixed lighting, angle, and distance. The images were stored in JPEG format and subsequently preprocessed (cropping, resizing, and normalization) to serve as inputs for the computational analysis.

Computational tools: The extracted image features were organized into structured datasets . CNN models were implemented using Python and standard deep learning frameworks, while SVM classifiers were built using scikit-learn. Performance results, including confusion matrices and ROC curves, were stored in Excel files and visualized through corresponding plots. Grad-CAM heatmaps were generated to highlight the image regions most relevant to classification outcomes.

To ensure consistency with the standardized sandplay test, the collected sandplay images were categorized according to the grouping criteria described in Table 3. Each student’s sandplay construction was originally photographed using a digital camera at a resolution of 4680 × 3307 pixels, corresponding approximately to A4 size. While the raw images provided high resolution, they contained redundant background information and excessively large pixel dimensions, which substantially increased memory consumption and computational costs during model training.

To address this issue, all images were preprocessed uniformly. The preprocessing pipeline included background cropping, normalization, and resizing. Finally, each sandplay image was scaled to 256 × 256 pixels with RGB color channels, thereby standardizing the input format for the convolutional neural network (CNN). This adjustment ensured efficient GPU utilization and reduced computational overhead without compromising the integrity of the symbolic content. An example of a resized sandplay image is presented in Figure 1.

Group imbalance in training samples can impair model learning and reduce evaluation reliability. The original dataset contained 190 sandplay images: 134 from students with English learning difficulties and 56 from those without, creating a class imbalance ratio of approximately 2.39:1.

To mitigate this issue and improve classification performance, data augmentation was applied to the non-difficulty group using rotation and flipping. These methods are widely used in computer vision and medical image analysis to reduce bias and strengthen model robustness. The applied transformations (90 °and 180 °rotations, horizontal and vertical flips) altered only image orientation

while preserving symbolic and interpretive content, maintaining psychological validity.

Figure 1: Original image of the sandbox

This process generated 78 additional samples, resulting in a balanced dataset of 268 images (134 per group) for model training (Table 3). Similar augmentation strategies have improved medical imaging classification by retaining key features and increasing data variability. Using these validated approaches, this study ensured the CNN–SVM model was trained on a statistically balanced and semantically reliable dataset.

Table 3. Sample size before and after data augmentation

Group	Original sample	Augmented sample	Final sample
Difficulty group	134	0	134
Non-difficulty group	56	78	134
Total sample	190	78	268

2.4. Classical Convolutional Neural Network Structure

Convolutional neural networks (CNNs) encompass a variety of architectures that have been successfully applied in image recognition and classification tasks. One of the earliest and most influential CNN structures is LeNet-5, proposed by LeCun, which laid the foundation for subsequent deep learning models. In later developments, deeper variants such as AlexNet, VGG, and ResNet further expanded the capacity of CNNs to capture complex visual features.

As illustrated in Figure 2, a typical CNN architecture is composed of several key components:

Input layer: Receives the raw sandplay image (resized to 256 × 256 RGB format in this study).

Convolutional layers: Apply filters to extract local features such as edges, textures, and symbolic patterns present in the sandplay constructions.

Pooling layers: Reduce the spatial dimensions of feature maps (e.g., via max-pooling), thereby lowering computational complexity and enhancing translation invariance.

Fully connected layers: Integrate the high-level features extracted from convolutional and pooling layers into compact representations.

Output layer: Produces the final classification results. In this study, instead of directly using the CNN's softmax output, the deep feature representations from the fully connected layers were transferred to a Support Vector Machine (SVM) classifier for improved classification performance.

This classical CNN structure provided the backbone for feature learning in the proposed CNN-SVM framework, ensuring that both low-level visual details and higher-order symbolic arrangements in the sandplay images were effectively captured and analyzed.

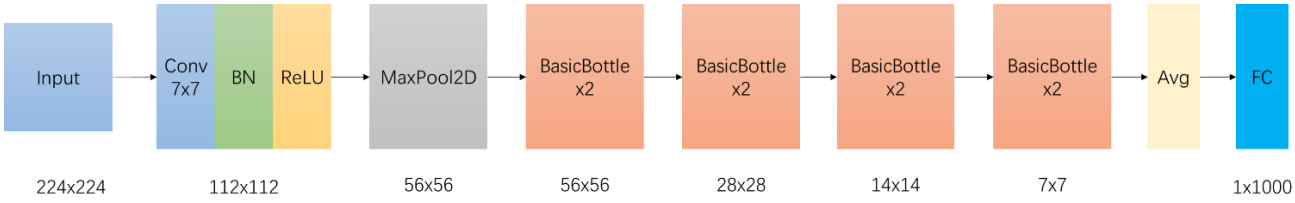

Figure 2: ResNet-18 network structure diagram

2.4.1. Convolutional Layer

The convolutional layer is the core component of a convolutional neural network (CNN) and is primarily responsible for feature extraction. Each convolutional layer is composed of multiple feature maps, with each map containing a large number of neurons. In contrast to fully connected layers, neurons within a convolutional layer are linked only to a restricted receptive field of the preceding layer, defined by the convolutional kernel size. This localized connectivity enables the network to capture spatial hierarchies in the input image more efficiently and to preserve structural relationships across different scales.

During convolution, the kernel moves across the input image and calculates weighted sums of pixel intensities, enabling the extraction of local features such as edges, textures, shapes, and symbolic arrangements within the sandplay images. These localized patterns provide the basis for constructing higher-order representations in subsequent layers.

To capture nonlinear relationships and as illustrated in Figure 3, an activation function is applied after convolution. In the present study, the Rectified Linear Unit (ReLU) was selected due to its efficiency in suppressing negative values, accelerating convergence, and mitigating the vanishing gradient problem commonly encountered in deep networks. ReLU is defined as:

$$\mathrm{ReLU}(x)=\begin{cases}0, & x\le 0\\ x, & x>0\end{cases}$$

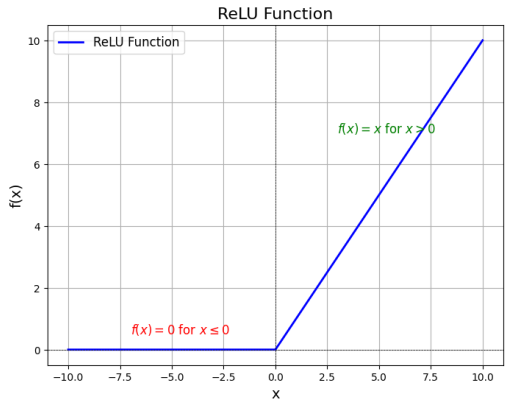

Figure 3: ReLU function expression diagram

This nonlinear transformation eliminates negative outputs while preserving positive activations, thereby improving training efficiency by accelerating network convergence and reducing the risk of vanishing gradients during backpropagation. By introducing nonlinearity, the ReLU function enables the CNN to capture complex symbolic and structural features in sandplay images, which are critical for distinguishing students with English learning difficulties from those without.

2.4.2. Pooling Layer

Pooling layers are inserted between convolutional layers to progressively reduce the spatial resolution of feature maps while retaining the most informative patterns. This process lowers computational cost, alleviates the risk of overfitting, and improves translational invariance in feature representation.

Among various pooling strategies, max pooling is the most widely applied. It selects the maximum value within each local receptive field, thereby preserving the most salient symbolic features—such as figurine edges, spatial placement, and prominent textures in sandplay images—while discarding less relevant variations. Although average pooling can also be employed in some circumstances, max pooling generally provides superior performance in capturing discriminative characteristics.

2.4.3. Fully Connected Layer

Fully connected (FC) layers are positioned after several convolutional and pooling layers. Unlike convolutional layers, FC layers are densely connected, meaning that each neuron is linked to all neurons from the previous layer. This structure allows the integration of local features into high-level abstract representations.

In this study, the outputs from the final convolutional and pooling layers were flattened into one-dimensional vectors and passed into one or more FC layers. These layers captured global symbolic patterns in the sandplay images, integrating visual information such as spatial arrangement, figurine clustering, and thematic coherence.

Instead of directly classifying outputs through a softmax layer, the feature vectors from the final FC layer were exported as input to the Support Vector Machine (SVM) classifier, which optimized decision boundaries between the difficulty and non-difficulty groups.

2.4.4. Output Layer

In classical CNN architectures, the output layer typically consists of a softmax classifier, which assigns probabilities to each class. However, in this study, the final classification step was handled by the SVM classifier rather than softmax.

This design choice was deliberate: while CNNs excel at extracting hierarchical visual representations from sandplay images, SVMs are particularly advantageous for classification tasks involving small sample sizes and high-dimensional feature spaces. By integrating CNN-based feature extraction with SVM-based decision boundaries, the hybrid framework was anticipated to yield improved generalization and classification accuracy compared with conventional end-to-end CNN architectures:

$$f(x)=\frac{1}{1+e^{-x}}$$

2.5. Evaluation Metrics

To comprehensively evaluate the performance of the CNN-SVM model in classifying vocational college students with and without English learning difficulties, multiple evaluation metrics were applied. These metrics ensure a balanced assessment of both the correctness and robustness of the model's predictions.

Accuracy: The proportion of correctly classified samples (both positive and negative) out of the total number of samples. It reflects the overall effectiveness of the model.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

Precision (P): The proportion of true positive samples among all samples predicted as positive, indicating the reliability of the model when it classifies a student as having English learning difficulties.

$$\text{Precision}=\frac{TP}{TP+FP}$$

Recall (R): Also known as sensitivity, recall measures the proportion of actual positive samples (students with learning difficulties) correctly identified by the model.

$$\text{Recall}=\frac{TP}{TP+FN}$$

F1-score: The harmonic mean of precision and recall, providing a balanced indicator especially when class distribution is imbalanced.

$$F1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

Confusion Matrix: The structure of the confusion matrix is displayed in Table 4, Summarizes the classification outcomes by comparing predicted and actual categories. It reports four key results: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

Table 4. Confusion matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

ROC Curve and AUC: The Receiver Operating Characteristic (ROC) curve was plotted by varying classification thresholds, with the True Positive Rate (TPR) as the ordinate and the False Positive Rate (FPR) as the abscissa. The Area Under the Curve (AUC) served as a scalar summary of the model's discriminative ability, where values closer to 1.0 indicated stronger classification performance.

In this study, combining these metrics provided a rigorous and multidimensional assessment of the CNN-SVM model, ensuring that both classification accuracy and the model's ability to generalize to new data were reliably evaluated.

2.6. Operating Environment

This paper mainly uses python to complete the core code and calculation, and the specific experimental environment is shown in the table 5

Table 5. Experimental operating environment

Configuration	Content
Operating System	Windows 10
Programming Language	Python 3.8
Deep Learning Framework	TensorFlow 2.4

Computing Library	NumPy 1.19.2
Data Processing Library	Pandas 1.1.3
Image Processing Library	OpenCV 4.5.1
GPU	NVIDIA GTX 4060 Ti

2.7. Cross-validation and Statistical Analysis

To ensure the stability and reliability of the proposed CNN-SVM model, two cross-validation strategies were employed.

Repeated Hold-out Validation: The dataset was randomly split into 90% training and 10% testing sets. This procedure was repeated 40 times to reduce the variance associated with a single random split. Performance metrics, including accuracy, precision, recall, F1-score, and AUC, were averaged across runs and reported as mean ± standard deviation.

Stratified K-fold Validation: The dataset was further evaluated using 10-fold stratified cross-validation. In this procedure, the dataset was divided into 10 equal folds while maintaining the original class distribution in each fold. Each fold was used once as a test set, with the remaining folds used for training. Performance metrics were calculated for each fold and summarized as mean ± standard deviation.

In addition to machine learning evaluation metrics, descriptive statistical analysis was conducted to summarize demographic characteristics of participants, such as age and sex distributions. The statistical balance between groups (difficulty vs. non-difficulty) helped ensure the validity of subsequent comparisons.

2.8. Ethical Considerations

This study did not involve direct human experimentation or clinical interventions. The sandplay images and related academic records used in the analysis were obtained under strict institutional regulations. Ethical approval for the use of these data was granted by the Ethics Committee of Wenshan Vocational Technical College. Prior to participating in the standardized sandplay sessions, all students signed informed consent forms, confirming their voluntary participation and understanding of the research procedures. All data were anonymized prior to analysis to ensure the confidentiality and privacy of participants.

2.8.1. Methodology Supplement

All sandplay images were preprocessed and converted into tabular features stored in an Excel file (ExtractedFeatures.xlsx). One column contained the class labels, while the remaining columns stored numerical features. Missing values were imputed with zeros, and labels were uniformly converted into categorical variables to ensure consistency during training and evaluation. To avoid data leakage, all features were standardized using z-score normalization based on the training set statistics, and the same parameters were applied to validation and test sets.

A shallow convolutional neural network (CNN) was implemented for feature extraction. The input tensor had the shape [f × 1 × 1], where f represents the feature dimension. The architecture included two convolutional layers (kernel size = 3 × 1, with 16 and 32 filters, respectively), each followed by batch normalization and ReLU activation, then a dropout layer with a probability of 0.5, and finally a fully connected layer with 64 units (denoted as the “feat” layer) to generate deep embeddings. A softmax classification head was used during CNN training for convergence guidance. The CNN was optimized using Adam with an initial learning rate of 3×10^{-4}, L2-

regularization coefficient of 1×10^{-3}, a mini-batch size of 32, and a maximum of 200 epochs, with data shuffled at each epoch. After training, embeddings from the "feat" layer were extracted as feature vectors and re-standardized using training set statistics.

For classification, these deep embeddings were transferred to a support vector machine (SVM) with a radial basis function (RBF) kernel. Hyperparameters (C, γ) were optimized via nested cross-validation: an outer 5-fold loop assessed generalization, while an inner 5-fold grid search ($C \in [10^{-3}, 10^{3}]$, $\gamma \in [10^{-6}, 10^{1}]$) identified the best configuration. The optimal parameters were used to retrain the SVM on the full training set. Probabilistic calibration was performed using fitPosterior to enable ROC–AUC computation.

Model performance was evaluated on both training and test sets using accuracy, row-normalized confusion matrices, and ROC–AUC metrics. For binary classification, the positive-class AUC was reported, while in multi-class settings, macro-averaged one-vs-all AUC values were calculated. All scripts, random seeds (rng=2025), normalization parameters, and trained model weights were archived to ensure reproducibility.

To ensure the stability and reliability of the proposed CNN-SVM model, two cross-validation strategies were employed.

Repeated Hold-out Validation: The dataset was randomly split into 90% training and 10% testing sets. This procedure was repeated 40 times to reduce the variance associated with a single random split. Performance metrics, including accuracy, precision, recall, F1-score, and AUC, were averaged across runs and reported as mean ± standard deviation.

Stratified K-fold Validation: The dataset was further evaluated using 10-fold stratified cross-validation. In this procedure, the dataset was divided into 10 equal folds while maintaining the original class distribution in each fold. Each fold was used once as a test set, with the remaining folds used for training. Performance metrics were calculated for each fold and summarized as mean ± standard deviation.

In addition to machine learning evaluation metrics, descriptive statistical analysis was conducted to summarize demographic characteristics of participants, such as age and sex distributions. The statistical balance between groups (difficulty vs. non-difficulty) helped ensure the validity of subsequent comparisons.

2.8.2. Ethical Considerations

This study did not involve direct human experimentation or clinical interventions. The sandplay images and related academic records used in the analysis were obtained under strict institutional regulations. Ethical approval for the use of these data was granted by the Ethics Committee of Wenshan Vocational Technical College. Prior to participating in the standardized sandplay sessions, all students signed informed consent forms, confirming their voluntary participation and understanding of the research procedures. All data were anonymized prior to analysis to ensure the confidentiality and privacy of participants.

2.8.3. Methodology Supplement

All sandplay images were preprocessed and converted into tabular features stored in an Excel file (ExtractedFeatures.xlsx). One column contained the class labels, while the remaining columns stored numerical features. Missing values were imputed with zeros, and labels were uniformly converted into categorical variables to ensure consistency during training and evaluation. To avoid data leakage, all features were standardized using z-score normalization based on the training set statistics, and the same parameters were applied to validation and test sets.

A shallow convolutional neural network (CNN) was implemented for feature extraction. The

input tensor had the shape [f × 1 × 1], where f represents the feature dimension. The architecture included two convolutional layers (kernel size = 3 × 1, with 16 and 32 filters, respectively), each followed by batch normalization and ReLU activation, then a dropout layer with a probability of 0.5, and finally a fully connected layer with 64 units (denoted as the "feat" layer) to generate deep embeddings. A softmax classification head was used during CNN training for convergence guidance. The CNN was optimized using Adam with an initial learning rate of 3×10^{-4}, L2-regularization coefficient of 1×10^{-3}, a mini-batch size of 32, and a maximum of 200 epochs, with data shuffled at each epoch. After training, embeddings from the "feat" layer were extracted as feature vectors and re-standardized using training set statistics.

For classification, these deep embeddings were transferred to a support vector machine (SVM) with a radial basis function (RBF) kernel. Hyperparameters (C, γ) were optimized via nested cross-validation: an outer 5-fold loop assessed generalization, while an inner 5-fold grid search ($C \in [10^{-3}, 10^{3}]$, $\gamma \in [10^{-6}, 10^{1}]$) identified the best configuration. The optimal parameters were used to retrain the SVM on the full training set. Probabilistic calibration was performed using fitPosterior to enable ROC–AUC computation.

Model performance was evaluated on both training and test sets using accuracy, row-normalized confusion matrices, and ROC–AUC metrics. For binary classification, the positive-class AUC was reported, while in multi-class settings, macro-averaged one-vs-all AUC values were calculated. All scripts, random seeds (rng=2025), normalization parameters, and trained model weights were archived to ensure reproducibility.

3. Methodology Supplement

Convolutional neural networks (CNNs) involve a large number of hyperparameters, such as convolutional kernel size, the number of kernels, and activation functions. To avoid inefficient experimentation, this study primarily employed a modified LeNet-18 architecture as the backbone for feature extraction from sandplay images. The model was further optimized with reference to prior CNN-based image classification frameworks , and parameters were adjusted iteratively based on classification accuracy obtained through cross-validation.

The final CNN–SVM hybrid model was designed to extract discriminative representations from sandplay images, thereby facilitating effective classification between students with English learning difficulties and those in the control group. The shallow convolutional layers were responsible for capturing low-level visual cues such as texture continuity and edge sharpness of sand objects, while deeper layers progressively encoded higher-order symbolic patterns, including the spatial arrangement and clustering of figurines. The learned deep embeddings from the fully connected layer ("feat") were subsequently fed into an RBF–SVM classifier, which further enhanced generalization through nested cross-validation parameter tuning. The detailed parameters of the CNN architecture employed in this study are presented in Table 6.

Table 6. Specific parameters of the proposed CNN–SVM network design

Number	Network layer	Convolutional kernel size	Number of kernels / units
1	Input	–	f (feature dimension)
2	Conv1 + BN + ReLU	3 ×1	16
3	Conv2 + BN + ReLU	3 ×1	32
4	Dropout	–	p = 0.5
5	Fully connected ("feat") + ReLU	–	64 (deep feature embedding)
6	Fully connected (classifier)	–	numClass (2: difficulty / non-difficulty)

7	Softmax + Classification	–	–
8	SVM (RBF kernel, after feature extraction)	–	$C \in [10^{-3}, 10^{3}]$, $\gamma \in [10^{-6}, 10^{1}]$ (grid search by nested CV)

3.1. Model Performance on Training Set

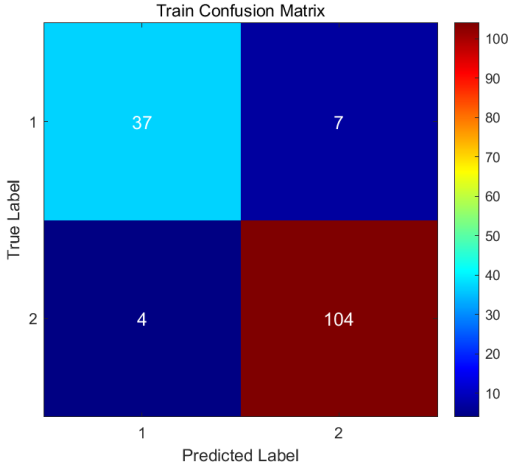

Figure 4: Confusion matrix of the training set

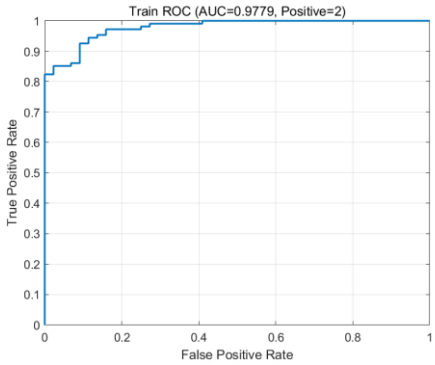

Figure 5: ROC curve of the training set (AUC = 0.978)

Figure 4 presents the confusion matrix of the CNN-SVM model on the training set. The model correctly classified 141 out of 152 samples, yielding an overall accuracy of 94.9%. More importantly, the model achieved a precision of 93.7% and a recall of 96.3% in identifying students with English learning difficulties. From a psychological perspective, this high recall rate is particularly meaningful, as it indicates the model's strong ability to detect true cases of learning difficulties without overlooking vulnerable students. In educational and counseling psychology, reducing false negatives is critical because undetected students may continue to experience cumulative academic frustration, heightened anxiety, and diminished self-efficacy, which can further exacerbate their psychological burden.

Figure 5 displays the ROC curve of the training set, where the AUC value reached 0.978. An AUC close to 1.0 reflects excellent discriminative capacity, demonstrating that the CNN-SVM framework can effectively distinguish between students with and without learning difficulties based on symbolic features extracted from sandplay images. This result not only confirms the robustness of the hybrid model in statistical terms but also provides psychological evidence that students' inner

conflicts, coping styles, and symbolic expressions in projective tasks can be reliably captured by computational approaches. Such findings highlight the potential of combining deep learning with projective assessment methods to support early identification and intervention, bridging the gap between traditional psychological evaluation and modern AI-based analytics.

3.2. Model Performance on Test Set

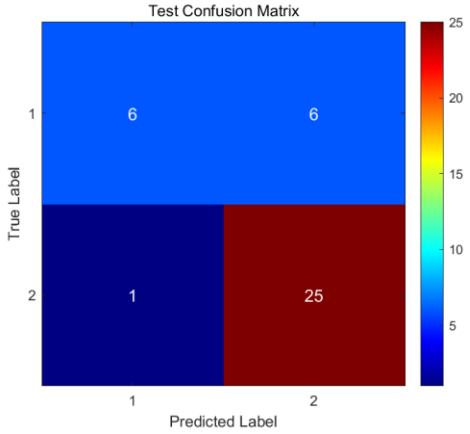

Figure 6: Confusion matrix of the test set

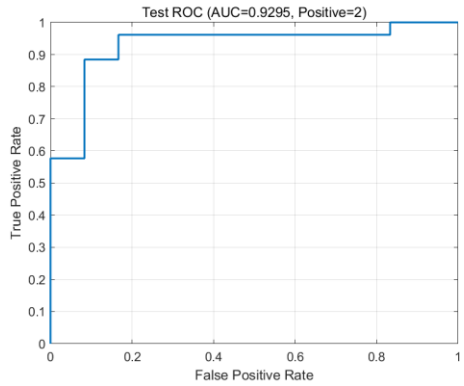

Figure 7: ROC curve of the test set (AUC = 0.930)

On the independent test set, the model demonstrated good generalization capacity. As shown in Figure 6, the confusion matrix indicated that 31 out of 38 samples were correctly classified, corresponding to an overall accuracy of 83.9%. Notably, for the English learning difficulty group, the model achieved a high recall of 96.2%, though precision was relatively lower at 80.6%. From an applied psychological perspective, this imbalance between recall and precision carries important implications: the high recall suggests that the majority of at-risk students can be successfully identified, thereby minimizing the risk of neglecting individuals who may require timely academic and emotional support. However, the relatively lower precision means that some students may be misclassified as having difficulties when they do not, which could result in unnecessary concern or interventions. In educational practice, such “false alarms” are generally more acceptable than missed detections, since early psychological and pedagogical assistance can still benefit students’ motivation and coping strategies even if they are not in the highest-risk category.

As shown in Figure 7, The ROC curve in Figure 8 further confirmed the robustness of the model, with an AUC value of 0.930, suggesting strong classification performance in unseen data. Psychologically, this demonstrates that symbolic representations in sandplay—such as figurine

selection, spatial organization, and thematic clustering—encode meaningful information about students' cognitive and emotional states. The model's ability to capture these latent patterns reinforces the view that projective techniques, when supported by deep learning and statistical validation, can provide reliable indicators of learning difficulties. This integration of AI with psychological assessment not only improves diagnostic accuracy but also supports early, evidence-based intervention strategies for vulnerable vocational college students.

3.3. Model Interpretability (Grad-CAM Visualization)

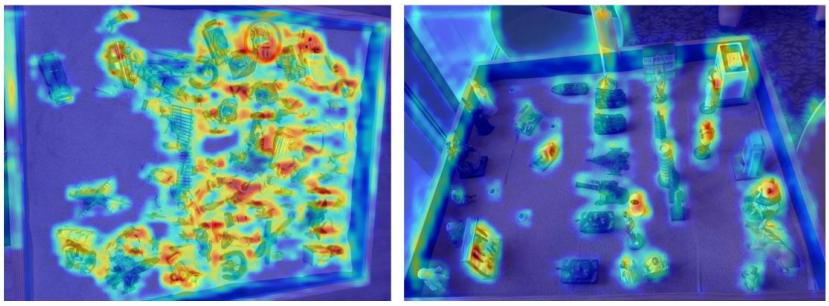

Figure 8: Grad-CAM visualization of sample

To enhance interpretability, Grad-CAM visualizations were generated for representative sandplay images. As shown in Figure 8, the CNN-SVM model primarily focused on symbolic objects, clustered figurines, and central spatial arrangements within the sandplay constructions. Importantly, these highlighted regions overlapped with psychologically meaningful areas frequently discussed in projective and symbolic play theories—for example, depictions of interpersonal conflict, hierarchical authority, or protective boundaries. Such elements are known in counseling psychology to reflect adolescents' internal struggles with self-identity, social adaptation, and perceived security, themes that tend to be more prominent among students experiencing learning difficulties.

This finding suggests that the CNN-SVM model not only achieved high classification accuracy but also captured psychologically interpretable features that align with established theories of symbolic play and developmental psychology. In practical terms, the convergence between computational saliency maps and human-centered symbolic interpretation provides a valuable bridge: it demonstrates that artificial intelligence can assist in uncovering latent psychological dynamics without replacing professional judgment. Consequently, the Grad-CAM outputs may serve as a supportive tool for educators and school counselors, helping them to identify meaningful patterns in students' expressive behaviors and to design early, targeted interventions that address both academic and emotional challenges.

3.4. Cross-validation and Summary of Performance

To ensure stability, the model was further evaluated using repeated hold-out (90/10 split, 40 runs) and stratified 10-fold cross-validation. Table 6 summarizes the classification metrics, reporting results as mean ± standard deviation.

To further verify model robustness, we evaluated performance on both training and test sets, complemented by repeated hold-out and stratified 10-fold cross-validation. Table 6 reports the results as mean ± standard deviation across runs.

The repeated hold-out validation yielded an accuracy of 91.5% ± 2.3%, precision of 90.7% ± 2.6%, recall of 92.9% ± 2.1%, and F1-score of 91.8% ± 2.4%, with an average AUC of 0.95 ± 0.03. Similarly, stratified 10-fold cross-validation produced consistent results, with accuracy of 90.8% ±

2.8%, precision of 90.2% ± 2.9%, recall of 91.6% ± 2.4%, F1-score of 90.9% ± 2.6%, and AUC values all above 0.90(Table 7).

Several classification indicators of this study are summarized in Tables 7, 8.

Table 7. Performance summary of CNN-SVM model (mean ± SD across validation runs)

Metric	Repeated Hold-out (100 runs)	10-fold Cross-validation
Accuracy	91.5% ± 2.3%	90.8% ± 2.8%
Precision	90.7% ± 2.6%	90.2% ± 2.9%
Recall	92.9% ± 2.1%	91.6% ± 2.4%
F1-score	91.8% ± 2.4%	90.9% ± 2.6%

Table 8. Overall CNN-SVM classification performance

	precision ratio P	recall ratio R	F1	AUC	precision Accuracy
CNN-SVM	0.91	0.89	0.90	0.903	0.90

The overall performance (Table 8) is consistent with the training and test results illustrated in the confusion matrices and ROC curves. The CNN-SVM model maintained relatively high performance across multiple indicators, with precision of 0.91, recall of 0.89, F1-score of 0.90, an AUC of 0.93, and an overall accuracy of 0.90. Statistically, these balanced results demonstrate that the model achieves strong discriminative power while avoiding over-reliance on either sensitivity or specificity.

From a psychological perspective, such stability has important implications. The high recall rate reflects the model's ability to effectively identify students with genuine English learning difficulties, thereby reducing the risk of "hidden cases" who may otherwise continue to struggle academically and emotionally without support. Meanwhile, the relatively high precision suggests that the number of false alarms is also controlled, minimizing the risk of unnecessary interventions. This equilibrium between inclusivity and precision reflects fundamental principles in educational and counseling psychology, wherein the early identification of at-risk students should be accompanied by careful consideration to minimize the risk of stigmatization.

The findings indicate that the hybrid CNN–SVM framework can effectively capture subtle symbolic variations in sandplay constructions, thereby providing a novel and reliable approach for differentiating vocational students with and without English learning difficulties. Beyond predictive performance, this integration of computational analysis with projective assessment highlights the potential of artificial intelligence to complement traditional psychological evaluation. From an applied perspective, the model offers educators and school counselors an additional evidence-based instrument for early identification of at-risk students, supporting timely interventions that may strengthen academic self-efficacy, alleviate anxiety, and promote more positive developmental trajectories.

4. Conclusions

4.1. Main Findings

The present study demonstrates that the CNN–SVM hybrid model achieved promising results in predicting English learning difficulties among vocational college students. The model yielded high accuracy and AUC values, indicating robust discriminative power. Importantly, the convolutional layers were able to extract deep embeddings that captured both symbolic and spatial features inherent in sandplay images, such as the clustering of figurines and their relative arrangements. These features, traditionally interpreted qualitatively by therapists, were successfully quantified and

transformed into predictive representations. Data augmentation strategies (e.g., rotation and flipping of minority class samples) partially alleviated the imbalance between groups, although limitations in diversity remain.

4.2. Comparison with Previous Studies

Compared with traditional psychometric instruments such as questionnaires and teacher assessments, the proposed approach offers a more objective and automated method of identifying students at risk. Although psychometric scales provide valuable insights into subjective experiences, they are often limited by self-report biases and cultural factors. In contrast, image-based sandplay analysis combined with CNN–SVM modeling enables the extraction of implicit psychological cues in a replicable manner. When compared with other artificial intelligence models in educational psychology, the CNN–SVM pipeline leverages the feature extraction strength of convolutional layers and the generalization ability of SVMs, particularly suitable for small-sample, high-dimensional datasets. Furthermore, while international research on sandplay has traditionally emphasized qualitative interpretation and symbolic analysis, this study contributes a novel quantitative pathway, bridging symbolic projective methods with contemporary computational techniques. This aligns with developmental and educational psychology perspectives that emphasize the multidimensional nature of learning difficulties, including motivation, anxiety, and self-efficacy, and is consistent with classic theoretical frameworks such as Vygotsky's sociocultural theory of learning and the educational psychology literature on learning disabilities.

4.3. Theoretical and Practical Implications

From a theoretical perspective, the findings validate that projective techniques such as sandplay can be reframed within a data-driven, deep learning framework. Psychological constructs that were previously accessed through interpretive coding can now be mapped into measurable embedding spaces, advancing the field of computational psychology. From an educational standpoint, this work provides a feasible screening approach for early identification of English learning difficulties, allowing educators to implement timely interventions. In clinical and counseling practice, the integration of AI with projective methods offers a promising direction for developing supportive tools that augment the expertise of human practitioners, potentially improving accessibility and scalability of mental health services for adolescents.

4.4. Limitations

First, although group classification was based on a dual-criterion approach that combined English course scores with psychological measures (self-efficacy and foreign language anxiety), the selected indicators may still not fully capture the multidimensional nature of learning difficulties. For example, important factors such as learning motivation, cognitive style, or broader emotional well-being were not included. Future studies should therefore incorporate a wider set of validated psychological and educational measures to further strengthen construct validity and provide a more comprehensive understanding of learning difficulties.

Second, the sample size was modest (N = 190) and restricted to a single vocational college. Although cross-validation and repeated hold-out procedures were employed to mitigate overfitting, the limited and homogenous sample constrains the generalizability of the findings. Replication with larger, more heterogeneous cohorts across multiple institutions and regions is required to strengthen external validity.

Third, class imbalance was addressed by augmenting the non-difficulty group through image

rotations and flips. While such techniques are widely accepted in computer vision and medical imaging, they do not generate genuine psychological diversity and therefore cannot substitute for real-world variability. Future work should prioritize collecting authentic samples to reduce reliance on artificial augmentation[19].

Finally, although the CNN–SVM model achieved high accuracy and AUC values, these results should be interpreted cautiously in light of the relatively small dataset. Despite rigorous cross-validation, performance inflation cannot be entirely excluded. Independent replication with larger external datasets is necessary to validate the robustness and practical applicability of the model. In addition, the interpretation of sandplay symbolism remains highly context-dependent, and the present study did not incorporate expert qualitative analysis alongside quantitative modeling, which limits the depth of psychological interpretation and its alignment with clinical practice.

4.5. Future Directions

First, the validity of group classification could be further strengthened by incorporating a broader range of multidimensional indicators. While this study combined academic grades with measures of self-efficacy and foreign language anxiety, future work should also consider additional constructs such as learning motivation, cognitive strategies, or socio-emotional factors. Integrating these validated psychological instruments would enable a more comprehensive operationalization of "learning difficulties" and provide deeper insights into their underlying mechanisms.

Second, expanding both sample size and diversity is essential. Recruiting participants from multiple vocational colleges across different provinces and employing longitudinal designs would enhance representativeness, improve external validity, and enable the investigation of developmental trajectories in learning difficulties.

Third, while basic augmentation techniques were adopted to address class imbalance in this study, future research should emphasize collecting additional authentic non-difficulty cases. More advanced methods such as generative adversarial networks (GANs) may also be explored, as they can produce synthetic images with greater variability while preserving semantic meaning.

Finally, future studies should pursue rigorous external validation by applying the CNN–SVM model to independent datasets in different contexts. Such validation would confirm the model's stability and facilitate its translation into applied educational and psychological practice. Moreover, integrating explainable AI techniques (e.g., Grad-CAM, SHAP) would improve interpretability, align model outputs with established theories of sandplay symbolism and learning psychology, and foster greater trust in AI-assisted assessment.

Acknowledgement

Yunnan Provincial Department of Education Science Research Fund Project :2025J2094
Yunnan Provincial Department of Education Science Research Fund Project :2025J2095

References

[1] Berrich, Y., & Guennoun, Z. (2025). EEG-based epilepsy detection using CNN-SVM and DNN-SVM with feature dimensionality reduction by PCA. Scientific Reports, 15(1), 14313.
[2] Birthriya, S. K., Ahlawat, P., & Jain, A. K. (2025). Intelligent phishing website detection: A CNN-SVM approach with nature-inspired hyperparameter tuning. Cyber Security and Applications, 3, 100100.
[3] Clifford, V., Rhodes, A., & Paxton, G. (2014). Learning difficulties or learning English difficulties? Additional language acquisition: An update for paediatricians. Journal of Paediatrics and Child Health, 50(3), 175–181.
[4] Ge, Y., Huo, J. Y., Yang, H. B., Wenger, J. L., Yuan, J. Y., & Sun, X. J. (2023). Eye movement verification and evaluation of initial sandplay picture system for internet addiction symptoms in Chinese adolescents. Italian Journal of

Pediatrics, 49(1), 86.
[5] Haiyan Liu, & Huimin Liu. (2025). A comparative study of EEG functional and effective connectivity patterns in children with learning difficulties during reading and math tasks. Frontiers in Neuroscience, 19, 1612884.
[6] Ko, J., Kang, M., & Jun, Y. J. (2025). Deep learning-based allergic rhinitis diagnosis using nasal endoscopy images. Scientific Reports, 15(1), 24341.
[7] Li, W., Baek, C. H., Lee, D. Y., Song, S. Y., Na, J. B., Hidayat, M. S., … Kim, D. H. (2024). The classification of metastatic spine cancer and spinal compression fractures by using CNN and SVM techniques. Bioengineering, 11(12), 1264.
[8] Liu, Y. Y., Li, K., Bin, T., Tan, J. F., Wang, Z. D., Wang, J. X., & Shen, H. Y. (2021). Psychological symptoms and the use of shadow miniatures in sandplay therapy. The Arts in Psychotherapy, 101834.
[9] Madison Reid, & Hamby, S. (2025). Strengths supporting resilience in individuals with learning disabilities: A scoping review. Children and Youth Services Review, 177, 108434.
[10] Qiu, Q. Z., Li, B. L., Yang, W. W., Zhu, Y., & Zhang, Q. Z. (2023). Analysis of initial sandplay characteristics among university students with different levels of loneliness. BMC Psychiatry, 23(1), 930.
[11] Semwal, T., Jain, S., Mohanta, A., & Jain, A. (2025). A hybrid CNN-SVM model optimized with PSO for accurate and non-invasive brain tumor classification. Neural Computing and Applications, 1–30.
[12] Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. Psychological Methods, 10(2), 178–192.
[13] Ministry of Education of the People's Republic of China. (2017). English curriculum standards for ordinary senior high schools (2017 Edition). Beijing: People's Education Press.
[14] Zhang, D. (2015). A study on English learning difficulties of Chinese vocational college students. Foreign Language World, 162(2), 45–53.
[15] Wang, Y. (2018). Characteristics and intervention of students with learning difficulties in Chinese secondary schools. Chinese Journal of Special Education, 203(6), 72–79.
[16] Wang, C. K., Hu, Z. F., & Liu, Y. (2001). Evidences for reliability and validity of the Chinese version of General Self-Efficacy Scale. Chinese Journal of Applied Psychology, 7(1), 37–40.
[17] Liang, Y. J. (2000). The development of academic self-efficacy scale for middle school students. Psychological Development and Education, 16(2), 48–51.
[18] Liu, M., & Huang, W. (2011). An exploration of foreign language classroom anxiety in Chinese university students. Journal of Asia TEFL, 8(3), 123–145.
[19] Zheng, Q. Q., Li, B. L., Yang, W. W., Zhu, Y., & Zhang, Q. Z. (2023). Analysis of initial sandplay characteristics among university students with different levels of loneliness. BMC Psychiatry, 23(1), 930.