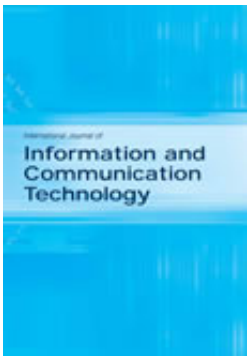


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Transfer learning-based adaptive music teaching system for modulating students' emotions

Mu Li

DOI: 10.1504/IJICT.2026.10075950

Article History:

Received:	22 October 2025
Last revised:	15 November 2025
Accepted:	17 November 2025
Published online:	06 February 2026

Transfer learning-based adaptive music teaching system for modulating students' emotions

Mu Li

College of Music and Communication,
Anyang University,
Anyang, 455000, China
Email: 001419@ayxy.edu.cn

Abstract: With the widespread application of artificial intelligence technology in the field of education, how to achieve emotion recognition and personalised regulation within the teaching process has become a significant research focus for intelligent teaching systems. This research provides an adaptive music teaching system solution based on transfer learning to address the constraints of traditional music education in emotional perception and delayed feedback. In this framework, learners' multimodal signals are initially acquired synchronously; later, transfer learning is utilised to facilitate emotion recognition and cross-domain feature transfer; ultimately, the system achieves synchronous optimisation of emotional regulation and learning performance. The proposed system shows improvements of 12.4%, 19.1%, 20.3%, and 17.4% in learning performance, engagement, emotional stability, and user satisfaction, respectively, when compared to traditional teaching techniques. This offers innovative theoretical frameworks and technological assistance for the development of emotion-driven intelligent educational systems.

Keywords: TL; adaptive music teaching system; emotion recognition; intelligent education.

Reference to this paper should be made as follows: Li, M. (2026) 'Transfer learning-based adaptive music teaching system for modulating students' emotions', *Int. J. Information and Communication Technology*, Vol. 27, No. 6, pp.66–89.

Biographical notes: Mu Li is a teacher in the College of Music and Communication at Anyang University, China. He received his Master's degree from Guangxi Arts University, China in 2012. He has published six papers and one monograph. His research interests include music education, music theory and emotion recognition.

1 Introduction

The rapid growth of artificial intelligence (AI) and educational technology is causing a big change in the field of education (Barakina et al., 2021). This change is based on data and backed by smart algorithms. Information technology is increasingly driving and changing music education, which is an important part of arts education. In the past, teachers' experience and the mood of the classroom were the main ways that traditional music lessons worked. Nonetheless, when faced with individual learner variances and

variable emotional conditions, it frequently encounters difficulties in attaining authentic personalisation and emotional adaptation. In modern higher education, the emotional and mental health challenges faced by university students are garnering heightened attention. How to use smart technology to improve the music education process, which will lead to better emotional experiences and more drive to study, has become a research area of great theoretical and practical importance.

Adaptive learning systems (ALS) have been used a lot in schools in the past few years. Their main idea is to change the content and ways of interacting with students based on their skill level, progress, interests, and emotional condition. This is how they achieve personalised training. Adaptive music teaching systems are a new type of smart teaching platform that was created based on this idea (Li and Han, 2023). These systems change their teaching methods, suggest topics, or change the settings for music generation in real-time based on how well students play music and what they say about it. This helps students get more involved and feel better about themselves while they are learning interactively.

Transfer learning (TL), a key machine learning (ML) technique, can solve the above problems. The introduction of TL can reduce these limits. By using models trained on large-scale emotion identification tasks to specific teaching scenarios, systems can accurately recognise and forecast emotions and states with minimal student data, supporting adaptive music instruction. Emotion recognition stands as a pivotal technology for human-machine affective interaction. Its objective is to discern and quantify learners' emotional states by analysing multimodal data encompassing physiological signals, vocal characteristics, facial expressions, or electroencephalogram (EEG) waves. In recent years, the convergence of affective computing with deep learning (DL) techniques has markedly enhanced both the accuracy and real-time capabilities of emotion recognition systems (Afzal et al., 2024).

With the proliferation of educational informatisation and AI technology, an increasing number of researchers are exploring the integration of emotion recognition technology into music education to develop intelligent systems capable of perceiving, understanding, and responding to students' emotional states. But there are still certain problems with existing adaptive music instruction technologies. First, the accuracy of emotion recognition is hampered by the lack of data sources and the fact that samples might vary; most systems only use vocal or facial expressions to figure out how someone feels, ignoring more consistent signs like EEG or physiological signals. Second, models do not generalise well, and their performance generally drops when they are used with students from different cultural backgrounds, musical styles, or learning environments (Ouyang et al., 2022). Thirdly, the emotional control mechanisms in the systems do not have dynamic feedback loops, which makes it impossible to have really adaptive interactions. Thus, employing TL to facilitate cross-domain knowledge transfer, augment the resilience of emotion recognition, and enhance the customisation of system responses represents the principal achievement of this research.

This research seeks to develop an adaptable music education system grounded in TL. The system uses an emotion detection module to detect how university students are feeling in real-time and then uses TL techniques to improve the model. This lets it keep its ability to recognise emotions with high accuracy even when there is not a lot of data in the target domain. When the system notices change in students' emotions, it automatically changes the content and speed of the lessons to improve both emotional management and learning outcomes. This study examines the adaptation of teaching

content and the evolution and feedback of learners' emotional states, investigating interactive mechanisms enabled by technological intervention, in contrast to previous studies.

This project aims to develop an emotion recognition model utilising deep neural networks (DNNs) at the technical implementation level. Using TL techniques, the model that was pre-trained on a generic emotion database will be used in a music teaching setting. By mapping features and fine-tuning parameters, the system will change and improve to fit different groups of learners. At the same time, the system design includes a closed-loop emotional feedback system that creates a dynamic cycle of interaction between changing the instructional content and detecting emotions. This makes a music teaching system that really adapts. Development will use Python and DL frameworks like TensorFlow. Real-time feedback and data visualisation will be possible through database and front-end interfaces.

This study presents several innovations: firstly, it proposes an emotion-driven adaptive teaching framework based on emotion recognition, thereby enriching the affective computing theoretical foundation of intelligent education systems; secondly, by integrating TL with emotion recognition methods, it enhances the model's performance across domains and in scenarios with limited data; thirdly, through empirical research in music teaching contexts, it reveals the mechanism by which emotional states influence learning experience optimisation and pedagogical interventions, providing reference for future development of emotion-aware learning systems.

From a practical standpoint, the results of this study can be immediately implemented in higher education, music paedagogy and mental health education. Adaptive music teaching systems allow teachers to keep an eye on how their students are feeling in real-time. This helps them improve their teaching methods and boosts their students' motivation to learn and their sense of self-efficacy through positive affect regulation. As AI and education continue to merge, these technologies are likely to become an important part of smart classrooms of the future, giving college students more customised and humanised learning experiences.

In summary, our project, located at the convergence of AI and educational technology, seeks to improve the emotional recognition and response functionalities of adaptive music teaching systems via TL technology. This investigation seeks to elucidate the prospective benefits of intelligent systems in emotional control and educational assistance. This project aims to establish the feasibility and application potential of emotion recognition in educational affective computing through methodical development, experimental validation, and data analysis. It offers both theoretical frameworks and practical instances for the development of a new generation of educational systems endowed with emotion-sensitive and individualised regulating functionalities.

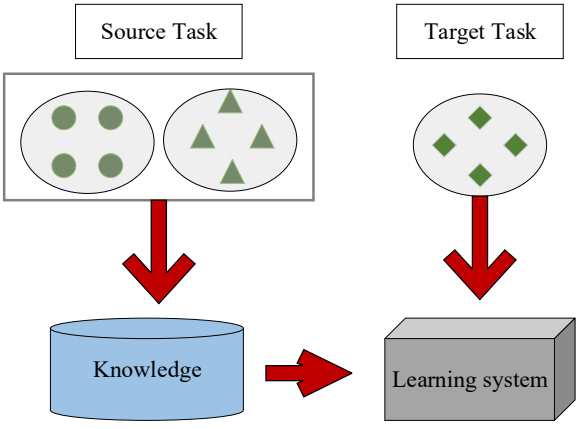
2 Relevant technologies

2.1 Transfer learning

TL is a major method in the field of ML that tries to use what is already known to improve how well people learn new tasks. Conventional ML models often presume that training and testing datasets derive from the same distribution and possess comparable task objectives. As a result, the performance of the model drops a lot when the data

distribution changes, the task changes, or there are not enough labelled examples. The fundamental tenet of TL is the transference of information from a source domain to a target domain (Zhang and Gao, 2022). This method lessens the requirement for labelled data in the target domain and improves the model's ability to adapt to new tasks or settings, as shown in Figure 1. The main goal of TL is to answer this question: how can information from existing models be efficiently reused to help learn new tasks when two tasks or data distributions are different but yet relevant?

Figure 1 The framework of TL (see online version for colours)



Formally defining the TL problem, let the source domain be D_S and the corresponding source task be D_S. Similarly, the target domain is D_T and the target task is T_T. When $D_S \neq D_T$ or $T_S \neq T_T$, the objective of TL is to optimise learning performance by leveraging knowledge from both D_S and T_S. Mathematically, this can be expressed as:

$$L_{TL} = \min_{f_T} \mathbb{E}_{(x,y)\sim D_T} \left[L\left(f_T(x), y\right) \right] + \lambda \Omega\left(f_T, f_S\right) \tag{1}$$

where $L(\cdot)$ is the loss function, $\Omega(f_T, f_S)$ is the term that shows how related the source and target tasks are, and λ is the balancing parameter. This formula shows the core of TL: it uses the regularisation term to transfer and modify source information effectively while minimising the loss of the target task.

TL can be divided into four main forms based on how it transfers knowledge: Inductive TL, transductive TL, unsupervised TL, and multi-task TL. Inductive TL deals with situations where the source and target tasks are different but have similar data distributions. This is usually used when the source task is labelled but the target task is not. Transductive TL deals with situations where the source and target tasks are the same but have different distributions, like when you need to adapt to a new domain. Unsupervised TL is mostly used for unlabelled target tasks, and it works by learning latent feature representations to share knowledge. Multi-task TL focuses on learning shared representations across multiple tasks at the same time to improve performance. These four categories each have their own use cases in different application environments, which is the theoretical basis for the TL methodology framework.

The main problem with TL when it comes to implementation strategies is figuring out what transferable knowledge is. There are four main types of TL approaches based on

how knowledge is stored: instance-based transfer, feature-based transfer, parameter-based transfer, and relation-based transfer. Instance-based approaches make things consistent by changing the weights or filtering the samples from the source domain so that they match the distributions of the target domain (Chen et al., 2022). Feature-based approaches employ feature space transformations or alignment techniques to alleviate distributional discrepancies between domains. Parameter-based methods entail the sharing or partial sharing of model parameters, such as the weights of neural network layers. Relation-based methods move structural relationships between samples or features from one task to another. These can be graph structures or semantic relationships.

Fine-tuning is one of the most common methods used in modern DL frameworks. The basic idea is to pre-train a DNN model on data from a source domain and then fine-tune it on data from a target domain to acquire some knowledge transfer. For example, ImageNet for image tasks or BERT models for natural language processing are big datasets that are often used for pre-training (Min et al., 2023). During the fine-tuning phase, some layers' parameters are frozen, and updates are only made to higher-level or particular modules. This makes it easy to quickly adapt to the target domain task. This method uses generalisable representations learnt from vast amounts of data while avoiding the extra work and risk of overfitting that comes with training models from scratch. Fine-tuning has shown to be very effective in many areas, including visual recognition, audio recognition, text classification, and others. It has become the typical way to reuse DL models across tasks.

Domain adaptation is an important part of TL. Its goal is to achieve high-performance generalisation in target domains where the objectives are the same, but the data distributions are different. Because the distributions between domains are different by nature, models that were trained on source domain data usually do not work as well in target domains. As a result, domain adaptation approaches make it easier to share information by reducing the differences in feature distribution between the source and target domains. Common methods include using maximum mean discrepancy (MMD) to align features and using adversarial learning to train domain discriminators (Jiang et al., 2025). Deep domain adaptation has become a popular area of research in the last few years. Domain-adversarial neural networks (DANN) are an example of a model that uses adversarial mechanisms to help models learn features that are not specific to a certain domain while still being able to tell the difference between tasks. This improves the model's ability to generalise to new domains.

Feature transfer concentrates on acquiring common feature representations across various tasks or data distributions. The basic idea behind it is that there is a hidden high-dimensional feature space where the data distributions of the source and target domains can be translated to similar structural forms. Feature transfer methods usually use techniques for extracting features and changing their locations, like linear projection, sparse coding, and autoencoders. These methods improve the similarity and alignability of domain-specific features by mapping or rebuilding input features. In the DL paradigm, convolutional neural networks (CNNs) and Transformer architecture inherently exhibit strong feature abstraction capabilities (Cong and Zhou, 2023). As a result, feature transfer using multi-task shared layers has become a common method. Additionally, some research employs statistical distance metrics, including Kullback-Leibler divergence or Wasserstein distance, to quantify and reduce the discrepancy between feature distributions of the source and target domains, hence augmenting transfer efficacy.

Sharing parts of model parameters across different tasks or domains is how parameter sharing strategies help people learn new things. For example, in multi-task learning, tasks generally share fundamental network designs to share features. Higher-level networks, on the other hand, are optimised for their own tasks. This method lets models pick up on common traits between related tasks, which improves performance as a whole. Regularisation-based transfer methods are another type of transfer method. They add limits to the parameters of the source model inside the loss function of the destination job. This lets the model keep some of what it learned from previous tasks while it learns new ones.

In summary, TL bridges current knowledge with novel tasks, giving theoretical foundations and technical solutions to data shortages, task dissimilarity, and distribution shifts. Its main benefit is allowing robots to quickly adapt to new environments and activities through prior experience, like humans, reaching higher-level intelligent learning.

2.2 Adaptive music teaching system

Adaptive music teaching systems represent a significant research direction at the intersection of educational technology and AI. Their goal is to tailor training and maximise learning efficiency by dynamically adjusting teaching tactics and content based on learners' differences, learning styles, and musical performance.

Early research into adaptive music teaching can be traced back to the computer assisted instruction phase of the 1980s. The system at that time was primarily constructed using expert systems and logical rules, employing pre-set knowledge bases and decision tree (DT) to achieve tiered delivery of teaching content (Thaher et al., 2021). These systems had basic adaptive capabilities, but they depended a lot on rules that were written by hand and expert knowledge. Their capacity for learning and generalisation was constrained, making it difficult to meet the varied demands of learners. At the dawn of the 21st century, researchers began to focus on how to utilise learning process data to construct more refined adaptive models. Some systems, for example, use Bayesian networks or hidden Markov model (HMM) to look at changes in how well learners do and then change the difficulty and substance of the lessons on the fly. Research during this phase substantially propelled personalised teaching, transitioning adaptive music training from experience-based to data-informed methodologies.

At the same time that multimedia technology and online education have improved, online music teaching platforms have slowly appeared, and adaptive mechanisms have been added to networked teaching settings. Researchers are investigating methods to implement customised music education in virtual learning environments. Some systems, for example, use musical instrument digital interface (MIDI) technology to record student performance data in real-time. They then use algorithms to evaluate the quality of the performance and give rapid feedback. Other studies use web services and database technology to make remote teaching support possible. This lets teachers change their teaching plans on the fly based on how well students are doing.

Multimodal interfaces that include video, gestures, and voice are slowly taking the role of traditional keyboard or touchscreen inputs. Systems can learn more about how children learn by using cameras and sensors to record their playing posture, facial emotions, or lip movements. For example, several research initiatives have created virtual

fingering assistance systems that use cameras to find and fix learners' hand movements in real-time (Wang et al., 2023).

In recent years, research emphasis has increasingly transitioned towards augmenting the intelligence and humanisation of adaptive systems. Some researchers have started to focus on the long-term growth of learners and the development of their overall skills. They have introduced ideas like learner modelling and learning route recommendation. For instance, generative adversarial networks (GANs) or transformer models are employed to generate teaching materials, accompaniment music and practice pieces, thereby enhancing the diversity and creativity of system content (Takale et al., 2024). Such research advances the evolution of adaptive music teaching systems from supplementary teaching tools to intelligent teaching partners.

Recent studies have also looked on system explainability and fairness in schooling. As algorithms get more complicated, the black-box nature of how systems make decisions has become more obvious. Researchers suggest incorporating explainable AI techniques to improve the clarity of recommendation and feedback systems. For example, showing feature weights or using rule extraction methods helps teachers and students understand how the system makes decisions, which builds trust and acceptance. To eliminate learning disparities caused by data biases in personalised systems, researchers have suggested model designs based on fair learning and privacy protection. This will make adaptive music teaching systems more in line with educational ethics and social values.

In summary, adaptive music teaching systems, as a significant innovation within educational technology, have evolved from their early stages of computer-assisted instruction to a new era of intelligent teaching that integrates AI, learning analytics, and multimodal interaction. This process will give music education more and more strong technology backing and theoretical underpinnings for new ideas.

2.3 Emotion recognition

Early emotion recognition systems predominantly employed traditional ML approaches based on rules and feature engineering. Typical algorithms include support vector machines (SVM), HMM, and DT. A shared attribute of these methodologies was their dependence on manually crafted features. For example, local binary patterns or scale-invariant feature transformations were used to recognise facial expressions, and mel-frequency cepstral ccoefficients (MFCC) or Pitch were used to recognise speech emotions (Zhu et al., 2024). and examining frequency-domain and time-domain characteristics in physiological signal recognition, including heart rate variability and EEG power spectral density. Even though these methods worked well in some early studies, they could not be used in a lot of different situations since they were not very expressive, and the data distributions were too complicated.

As DL became more popular, emotion recognition technologies became mostly based on data. DNNs use multi-layered nonlinear transformations to automatically extract high-level semantic data. This greatly improves the accuracy and reliability of emotion recognition. CNNs are great at recognising emotions in pictures because they can automatically pick up on spatial elements in facial images, like muscle movements and micro-expressions. Recurrent neural networks (RNNs) and their variants, long short-term memory (LSTM) networks, are extensively utilised in speech and temporal emotion recognition, effectively capturing the dynamic attributes of emotional signals over time.

In recent years, models that use attention mechanisms and Transformer structures have made emotion recognition even better.

In the field of facial expression recognition, research has expanded from two-dimensional static images to three-dimensional dynamic sequences. Early systems primarily relied on the facial action coding system (FACS) rules to encode facial action units, whereas modern systems utilise deep convolutional networks (DCNs) to directly learn spatial-temporal features from video sequences, enabling the recognition and estimation of continuous emotions (Karnati et al., 2023). For example, models like 3D-CNN, C3D, and ConvLSTM can record both spatial texture changes and temporal dynamic information at the same time. This makes it much easier to recognise complicated emotions. This series of technological advancements has progressively shifted visual emotion recognition from discrete classification towards continuous prediction and multidimensional modelling.

Physiological signals are generally generated by autonomic nervous system reactions, rendering them less vulnerable to subjective manipulation and so enabling them to more accurately represent genuine emotional experiences. EEG signals, due to their great temporal resolution, are extensively utilised for real-time emotion detection and brain state analysis. Studies show that there are strong links between distinct frequency bands of brainwaves and emotional states. DCNs and Sequence Networks have shown excellent results in recognising emotions from EEG signals in recent years. They automatically extract complicated spatial and temporal information from multi-channel inputs.

Furthermore, emotion recognition research confronts many obstacles. First, data diversity and generalisability are issues: emotional expression varies greatly by culture, gender, and age. Distribution disparities between datasets sometimes hinder model transferability. Second, emotions are ambiguous and changing, therefore they rarely fit into stable categories. The move from discrete recognition to continuous estimation is an important research area. Thirdly, annotation and interpretability are difficult since identifying emotions is subjective and expensive, and deep models are hard to explain in real-world situations because they are black boxes (Hassija et al., 2024). Researchers have suggested methods like self-supervised learning, meta-learning, explainable models, and cross-cultural emotion recognition to make systems stronger, more open, and more universal in order to deal with these problems.

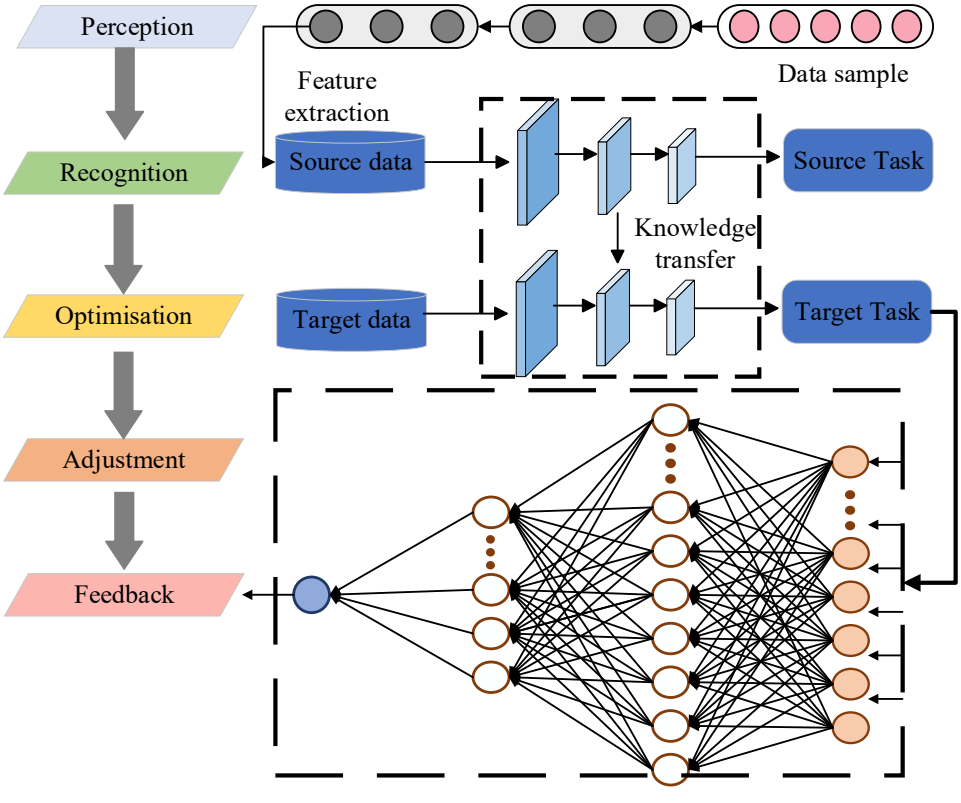
Emotion recognition technology has progressed from conventional feature-based approaches to DL-based intelligent modelling and is currently at the stage of multimodal and explainable fusion. As sensing technologies, computing power, and large-scale models continue to improve, emotion detection systems of the future will be more accurate, work in real-time, and work across different fields. They are ready to provide more value in education, healthcare, entertainment, and mental wellness.

3 System design and implementation

The TL-based adaptive music education system is built on a closed-loop logic of perception, recognition, optimisation, adjustment, and feedback. As shown in Figure 2, the system's goal is to make emotion recognition and instructional content work together in a dynamic way through multi-layered module collaboration. The system is made up of five main modules: data gathering and pre-processing, emotion identification and modelling, TL optimisation, adaptive teaching material adjustment, and system interface

and visualisation. From the bottom up, these five modules make up the system's input layer, analysis layer, intelligent optimisation layer, instructional regulation layer, and interactive feedback layer. They work together to let the system see and understand how learners are feeling in real-time, so it can provide them with tailored replies to their questions.

Figure 2 Architecture of the adaptive music teaching system (see online version for colours)



3.1 Data acquisition and preprocessing module

The module is the main part of the whole TL-based adaptive music instruction system. Its main job is to see, get, and standardise the data needed for recognising emotions and giving feedback on lessons. The way this module is built directly affects the quality of training for later models and the performance of TL. It is a key part of making sure the system is stable and accurate. This study utilises a multimodal fusion acquisition and a unified feature preprocessing technique. Data from different sources is mapped into a common feature space to meet the input needs of the upper-layer DNN (Liao et al., 2023). This is done using a standardised signal collection interface and feature engineering procedure.

The system uses multimodal sensing units to collect emotion-related signals and interaction behaviour data from learners while they are learning music. This module simultaneously acquires facial expression sequences from video monitoring, speech

signals from audio capture, and EEG data from physiological sensors. During collection, the system uses a unified sample frequency synchronisation method to make sure that multi-channel data is aligned in time. Drivers convert each sort of signal into digital form and store it in an organised database (Kostakis and Kargas, 2021). This makes sure that the data is always accurate and can be processed in real-time. All signals are sent to the preprocessing pipeline in fixed-window segments so that they may be processed in batches and the model can be trained consistently later.

The goal of data preprocessing is to get rid of noise, scale changes, and temporal drift from the raw signals. This makes feature representation more stable. First, all input signals are denoised in the frequency domain using bandpass filters to get rid of low-frequency baseline drift and high-frequency interference. Let $x(t)$ be the original signal and $H(f)$ be the filter transfer function. The filtered signal can be written as:

$$y(t) = F^{-1}\left\{H(f) \cdot F\left\{x(t)\right\}\right\} \tag{2}$$

where F and F^{-1} are the Fourier transform and inverse Fourier transform operations, respectively. This method makes sure that the signal is clear in the effective frequency band, which gives a clean database for further feature extraction. For each feature dimension xi, its standardised representation is:

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \tag{3}$$

where μ_i is the average of feature x_i, and σ_i is its standard deviation. After standardisation, all features follow a distribution with a mean of 0 and a variance of 1.

The data gathering and preprocessing module is the main part of emotion recognition and TL optimisation. Its output directly affects how accurate and adaptable the next model inference is. Overall, this module establishes a robust technical foundation for the system's emotion recognition, model optimisation, and instructional adaptation through a stable, scalable multimodal data acquisition and standardised preprocessing workflow.

3.2 Emotion recognition and modelling module

This module is the main part of the TL-based adaptive music teaching system. Its job is to accurately extract emotional characteristics from multimodal input data and transfer them to the right feature domain. It not only recognises the emotional states of learners in real-time, but it also transfers knowledge from the source domain (general emotion datasets) to the target domain (music learning situations). This keeps the recognition accuracy high even when there are not many target samples.

The system uses a DNN-based emotion recognition framework for its model architecture. A hierarchical structure that combines convolutional and fully linked layers creates a high-dimensional abstract representation of multimodal information. The input layer gets the fused tensor X output from the module that collects and cleans up the data. The convolutional layer uses parameter sharing to automatically find local dependency features in the time series. It then outputs a feature mapping matrix that is triggered by ReLU (Lara-Benítez et al., 2020). After that, this goes into a fully connected layer to create high-level emotional feature variables h. A Softmax function is used in the model's output layer to make probabilistic predictions about different emotional states. This may be written mathematically as:

$$P\left(y_i \middle| X\right) = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}} \tag{4}$$

where z_i stands for the linear output that goes with the i^{th} emotion category, and C stands for the total number of emotion categorisation categories. This function allows the model to output the probability distribution for each emotion category, which makes it possible to recognise the learner's emotional state in more than one way.

This work implements a TL technique predicated on feature alignment inside this module to facilitate model transfer between domains. The system initially engages in pre-training on a comprehensive emotion dataset (source domain) to acquire universal emotion feature representations. After that, the feature mapping function $\phi(\cdot)$ projects data from the source domain and target domain onto a shared feature space. This makes sure that the feature distributions are the same in both domains. Let $P_s(X_s)$ be the distribution of features in the source domain and $P_t(X_t)$ be the distribution of features in the target domain. The goal of feature transfer is to make the difference in distribution between the two as small as possible (Hosna et al., 2022). This goal for optimisation can be put into writing as:

$$L_O = \min_{\phi} D\left(\phi\left(X_s\right), \phi\left(X_t\right)\right) \tag{5}$$

where $D(\cdot,\cdot)$ is the function that measures the distance between feature distributions. This research utilises MMD as the optimisation objective for distribution distance, quantifying the similarity between distributions across domains via kernel function mapping in high-dimensional feature space, thus facilitating domain adaptation at the feature level.

The system uses a two-stage optimisation process during model training. The first step is source domain pre-training, in which the model learns from a large-scale emotion identification dataset with supervision so that it can completely understand basic emotional patterns. The second step is adapting to the target domain by fine-tuning on a modest amount of music teaching scenario data. During fine-tuning, only the weight parameters close to the output layer are changed (Ding et al., 2023). This keeps the general feature extraction ability of the layers before it from overfitting. The optimisation goal function is characterised as a weighted amalgamation of cross-entropy loss and distribution alignment loss:

$$L = L_{ce} + \lambda L_{mmd} \tag{6}$$

where L_{ce} is the loss for emotion recognition, L_{mmd} is the loss for aligning feature distributions, and λ is the balancing coefficient. The model improves cross-domain recognition performance by minimising classification error and domain distribution divergence at the same time through combined optimisation.

The system uses the TensorFlow framework at the network implementation level to make automatic gradient backpropagation and dynamic graph optimisation easier. To make training more stable, batch normalisation and Dropout regularisation techniques are used. These stop gradient explosions and overfitting. During model training, an early stopping mechanism is used. This means that when the validation set loss stops going down over several iterations, training stops automatically, and the best weight parameters are saved (Bai et al., 2021). This makes the model better at generalising.

The emotion recognition and feature transfer module work together to make the system work. The TL method lets you map low-level features into understandable emotional states at the same time as it lets you quickly share knowledge and change the model to fit new situations. The module's design takes full advantage of the benefits of DL and TL working together. This lets the system keep high accuracy and stability in recognising emotions even when there is not much data and people are very different from each other. This makes for a strong algorithmic base for further changes to teaching content and emotional feedback loops.

3.3 TL optimisation module

The TL optimisation module is the algorithmic heart of the TL-based adaptive music education system. It is in charge of cross-domain feature adaption and performance optimisation during model transfer. This module lets source domain pre-trained models converge quickly and keep a high recognition accuracy with less target domain data by adding parameter fine-tuning and feature mapping optimisation techniques. Its main goal is to fix performance problems that happen when there are differences in how data is distributed across the source and target domains. This will make emotion detection models in music education more stable and able to generalise.

The main parts of the workflow are feature layer alignment and parameter layer optimisation. First, the system reconstructs and normalises features taken from the source domain model to make sure that input features are statistically consistent across domains. Let $P_s(X_s)$ be the distribution of features in the source domain and $P_t(X_t)$ be the distribution of features in the target domain. The goal of optimisation is to make the difference between these distributions in the embedding space as small as possible. To align the domains, we use MMD to measure the distance between the two domains' distributions. This is technically written as:

$$L_{mmd} = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \phi\left(x_i^s\right) - \frac{1}{n_t} \sum_{j=1}^{n_t} \phi\left(x_j^t\right) \right\|^2 \tag{7}$$

where $\phi(\cdot)$ is the operator that maps the high-dimensional kernel function. By reducing L_{mmd}, the model efficiently eradicates differences at the feature level caused by domain bias, facilitating the robust transfer of emotional traits across varied circumstances (Yan et al., 2025).

The parameter-level fine-tuning layer is the most important part of the TL optimisation module. The model's overall optimisation objective function is defined as:

$$L_{total} = L_{ce} + \lambda_1 L_{mmd} + \lambda_2 L_{reg} \tag{8}$$

where L_{ce} stands for the cross-entropy loss for classification error, L_{reg} stands for the weight regularisation term to stop overfitting, λ_1 and λ_2 are the adjustment coefficients. This combined optimisation technique reduces classification error while making sure that feature distribution is aligned and model parameters may be updated smoothly.

The system has an adaptive learning rate scheduling technique that makes optimisation faster and more stable during the TL process. This system changes the learning rate in real-time based on how quickly the loss function is going down. It keeps greater strides during the early convergence phase to speed up learning and then

gradually reduces strides near the optimum to stop oscillations. At the same time, Dropout and Batch Normalisation are added to the module to stop overfitting and speed up training convergence, which will help the model generalise better.

The TL optimisation module can also optimise online, which means that it can make small changes to the system while it is running depending on fresh target domain data. When the system sees big changes in the distribution of input data, it uses the KL divergence to figure out how much the distribution has drifted. This technique updates local parameters with tiny batches of fresh samples, so the model does not have to be retrained from scratch.

3.4 Adaptive teaching content adjustment module

The module is the part of the TL-based adaptive music teaching system that makes decisions and lets people interact. It is in charge of turning results from emotional recognition and learner conduct into flexible teaching tactics, which allows for individualised presentation of instructional content and control over the pace of the lesson. This module changes the complexity of the instructional content, the choice of musical material, and the pace settings in real-time based on the learner's emotional state. This lets the system improve the learning experience and cognitive performance in different emotional states.

It has three parts: the emotional state analysis layer, the instructional strategy generation layer, and the instructional content execution layer. The emotional state analysis layer accepts the emotional variable E that the emotion recognition module outputs as input. It then uses an emotion evaluation model to show the learner's psychological condition in numbers. The system uses a two-dimensional emotional model for mapping. Valence shows whether an emotion is positive or negative, and Arousal shows how active an emotion is (Petrolini and Viola, 2020). Let the detected emotional attribute be designated as ei; the emotional state can then be represented as a two-dimensional variable:

$$S = (v, a) \tag{9}$$

where v stands for pleasure and a stands for arousal. The system fully balances emotional control and learning effectiveness using a weighted multi-objective optimisation function, which is defined as follows:

$$L = \max_{T} U(T) = \alpha \cdot L(T, B) + \beta \cdot R(T, S) \tag{10}$$

where $L(T, B)$ is the function that shows how the instructional task helps with learning behaviour, $R(T, S)$ is the function that shows how the instructional approach controls emotional state, α and β are the equilibrium coefficients. By optimising the utility function U(T), the system attains a dynamic equilibrium between emotional regulation and learning performance, thus fulfilling the educational aim of synergistic optimisation.

The system uses a two-layer decision-making process that combines rule-based and learning-based methods when it comes up with instructional strategies. The first layer has an emotional rule engine that uses pre-set behavioural mapping rules to respond quickly. For example, if the system sees that a student is in a low-arousal, low-pleasure state, it immediately makes the lessons easier and plays happy music in the background to make the student feel better. On the other hand, when students are in a state of high arousal and

high tension, the system slows down the speed and lowers the number of tasks to avoid cognitive overload. A strategy optimiser that uses reinforcement learning makes up the second layer. This optimiser keeps changing the weights of strategies depending on past interactions and emotional input, which makes training more personalised.

The instructional content execution layer takes the parameterised strategies made by the strategy layer and turns them into precise directives for how to behave in the system's multimedia teaching components. Changes to instructional content mostly show up in three ways: controlling the difficulty, controlling the pace, and choosing the musical material. In terms of difficulty modification, the system changes the complexity of the exercises and the frequency of feedback prompts in real-time to match the learner's present cognitive load (Seyderhelm and Blackmore, 2023). The module uses real-time parameter control algorithms to change the tempo and density of music playback to keep learners focused and comfortable. The system picks the right musical styles and melodic structures from its database based on the user's emotional state and the goals of the activity.

The system uses database interfaces to get learning records and emotional feature data in real-time. It then uses a RESTful API to connect to the front-end teaching platform and show audio, graphics, and text in real-time. The instructional content execution layer makes it possible to implement TensorFlow Serving, which makes sure that strategy inference and feedback responses happen quickly.

The adaptive teaching content adjustment module uses an emotion-driven intelligent decision-making mechanism to deeply connect instructional materials with the psychological states of learners. This module not only dynamically optimises the educational experience across diverse emotional states but also constantly learns and self-evolves through sustained interaction, gradually building individualised learning pathways. It therefore offers a novel framework for human-machine collaborative learning in higher education settings.

3.5 System interaction and visualisation module

The system interaction and visualisation module is the main user interface for the TL-based adaptive music education system. It is responsible for important functions including showing statistics, giving interactive feedback, and making the system easier to understand. The goal of its design is to make it easier for learners, teachers, and the system to share information quickly and easily. This will allow for the intuitive and dynamic visualisation of emotion recognition findings, instructional material adjustment procedures, and assessments of learning efficacy. This module not only improves how people and computers interact with each other, but it also makes sure that teaching data can be understood, and decisions can be made in a clear way. This makes sure that the whole adaptive teaching process can be tracked, understood, and improved.

The data visualisation layer is the main part of this module for analysing and presenting data. It does multidimensional visualisations for emotional data, learning performance measurements, and instructional approach parameters. The system uses multimodal data fusion techniques to combine the results of the emotion detection and instructional adjustment modules to construct multi-tiered visualisation models. If E stands for the emotional state set and P stands for the learning performance indicators, the system creates a composite performance index using a weighted fusion function:

$$I = \sum_{i=1}^{k} \omega_i e_i + \sum_{j=1}^{m} \gamma_j p_j \tag{11}$$

where ω_i and γ_j are the weighting coefficients for emotional and learning traits, respectively. This composite indicator gives a clear picture of the learner's emotional stability and learning efficiency. It can be shown in the interface in a number of different ways. In this way, the method not only helps students keep track of changes in their emotions and performance, but it also gives teachers measurable reasons to step in and help.

The feedback management layer enables real-time closed-loop interaction between learners and the system, based on a human-machine co-adaptive feedback paradigm. The system's operating logic is as follows: after implementing pedagogical adjustment procedures, it collects learner reaction signals through sensors and behavioural monitoring modules. To find the feedback error ε, these are compared to the desired targets:

$$\varepsilon = \left| S_t - S_{t+1} \right| \tag{12}$$

where S_t and S_{t+1} are the emotional state variables that show how the person feels before and after the change in instruction. If the feedback error is greater than the defined threshold δ, the system immediately starts a policy re-optimisation method to make small changes to the teaching parameters. This method allows the system to adaptively track and optimise learner conduct throughout ongoing engagement, creating a dynamic process of human-machine co-evolution.

This module uses a front-end architecture that combines python with React. The back end uses the Flask framework to construct API services that analyse input and send model inference results in real-time. The front-end uses React and D3.js to show data in a way that lets users interact with it. The system uses the WebSocket protocol to sync emotion detection results and changes to instructional content within milliseconds (Jagtap et al., 2023). This makes sure that both emotional changes and visualised displays happen right away. The module also uses caching and asynchronous rendering to make sure that the interface is stable and smooth, so that users may have smooth experience even when there are a lot of people using it at once.

The interaction and visualisation module also keeps track of data and records learning logs. All interaction events and emotional change data are stored in the database so that we may later analyse learning behaviour and improve teaching strategies. By looking at the logs, the system can find long-term patterns of emotional changes and learning problems. This sets up a system for ongoing improvement based on data.

The module not only shows emotional recognition results, teaching strategies, and learning performance in a multidimensional, dynamic way, but it also makes the system's intelligent interaction capabilities better by providing real-time feedback and making the data easier to understand. The architecture of this module changes how people and machines work together from one-way information sharing to a two-way, intelligent, co-adaptive process. This gives important technical support for making future intelligent education systems more understandable, clear, and trustworthy for users.

4 Experiment and results analysis

4.1 Experimental data

The experimental data utilised in this study is derived from two principal datasets: the source domain dataset and the target domain dataset. The primary aim of data production is to provide a multimodal, high-quality data foundation for the training, transfer, and validation of TL models. The two datasets differ in where they came from and what tasks they were made for, but they are structurally consistent in design. This makes it possible to recognise emotions across domains and improve instruction in a way that works best for each student.

The source domain dataset is mostly used to train the TL model before it is used for anything else. This dataset combines publicly available multimodal emotion recognition resources, such as FER2013 (facial expression data) and RAVDESS (speech emotion data), with ethically approved classroom emotion video samples from colleges and universities to improve the model's ability to recognise emotions in general. This dataset consists of over 28,000 multimodal samples, each having image frames, audio signals, and emotion labels (Valence, Arousal, Category) that were all recorded at the same time. Standardisation and label harmonisation were applied to all data to create a single feature space for feature transfer and parameter fine-tuning.

Table 1 Datasets characteristic information

Dataset	*Data type*	*Main indicators*	*Data source*	*Acquisition method*	*Description*
Source domain dataset	Emotional data	Facial expressions, voice tone, valence, arousal, emotion category	Public datasets (FER2013, RAVDESS)	Pre-collected and standardised multimodal data	Used for TL model pretraining to capture generalised emotional features
Target domain dataset	Emotional data, learning behavioural data, system interaction data	HRV, EDA, RR, Task completion rate, error rate, response time	University music learning scenarios	Real-time synchronous acquisition via sensors and system logs	Used for model fine-tuning and adaptive learning optimisation
Dataset	Data type	Main indicators	Data source	Acquisition method	Description

The second dataset, which comes from real-world data collected during music lessons at colleges and universities, is meant to show how learners' emotions and behaviour change when they interact with the system. The participants consisted of 80 individuals from three comprehensive colleges, exhibiting an almost 1:1 male-to-female ratio, aged 18 to 23 years, and included both music and non-music majors. The target domain dataset consists of three key types of data: emotional data, learning behaviour data, and system interaction data. The system captured all the data in real-time during training sessions, including heart rate variability (HRV), electrodermal activity (EDA), respiration rate

(RR), facial features, voice characteristics, and interaction logs. This produced almost 9,600 synchronised multimodal samples, offering genuine scenario assistance for TL model transfer training and validation.

Table 1 shows the main features and organisation of the datasets.

During the data preprocessing phase, data from both the source and target domains underwent multimodal synchronisation, temporal alignment, and feature normalisation. Butterworth filters were used to remove noise from physiological inputs, while face landmark identification and pose calibration were used to keep image data consistent. MFCC conversion was used to get acoustic information from speech data. In the end, all features were brought down to the [0, 1] range to get rid of disparities in dimensions between modalities.

To sum up, the experimental data framework created in this study creates a structural match and feature complementarity between the source and target domains. This dual-dataset architecture offers a solid data basis for the cross-domain adaptation of TL models and the validation of system adaptive performance.

4.2 Emotion recognition performance comparison experiment

The goal of this part is to test how well TL-based emotion identification models work with diverse types of data. Two datasets are used for the experiments: a source domain dataset and a target domain dataset. Experiments in the source domain mostly test the model's capacity to extract features and recognise emotions in general. Experiments in the target domain, on the other hand, mainly test the model's ability to recognise music instruction scenarios accurately and consistently. By comparing performance indicators across several DL and classical models, this study thoroughly checks the TL approach's usefulness and benefits across many domains.

This research evaluated seven emotion identification models across several paradigms, including classical ML, DNN, and TL architecture. The SVM model used a standard classification structure with an RBF kernel function as a starting point for basic performance testing. The RF model uses multi-DT integration to discriminate features, is a benchmark for performance, the CNN model captures local spatial features in multimodal data like images and speech, and the LSTM model can model temporal data, making it suitable for analysing dynamic emotional signal changes. The ResNet-50 model, which is a pre-trained convolutional network, is used for deep feature transfer and expression optimisation. The DANN model uses an adversarial domain adaptation mechanism to improve cross-domain recognition capabilities through feature distribution alignment.

We did pre-training tests on the source domains (FER2013 + RAVDESS) to see how well each model did on generic emotion recognition tasks. The evaluation parameters comprised accuracy and macro-average F1-score. Figure 3 shows the outcomes of the experiments:

We tested and transferred models in the target domain (music instruction scenario data) to see how well each model could adapt to small sample sets and changes in environment. All models, except for the proposed system, were trained or fine-tuned directly on the target domain while keeping the data partitioning the same. Figure 4 shows the outcomes of the experiment:

Figure 3 Performance comparison of emotion recognition models on the source domain dataset (see online version for colours)

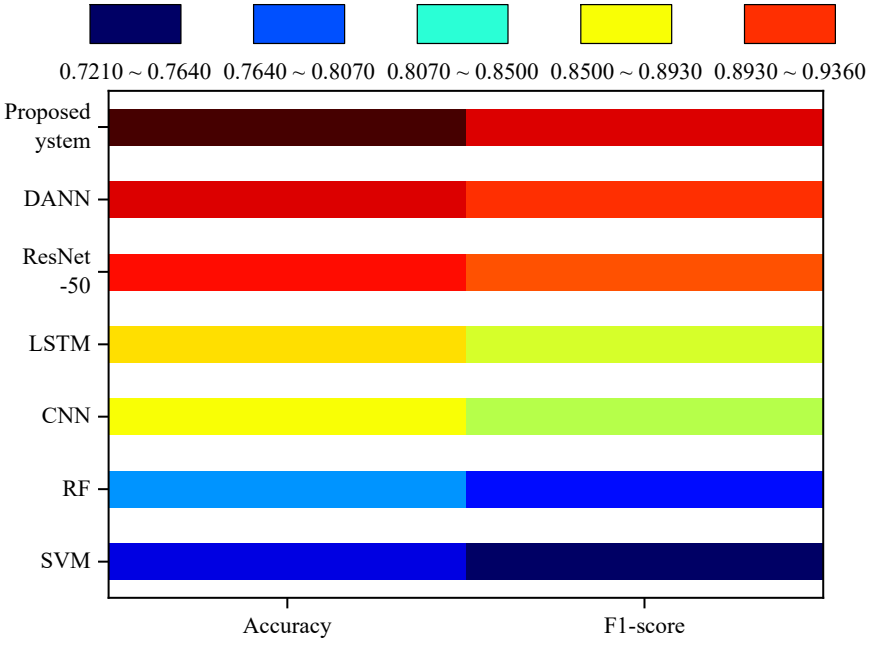
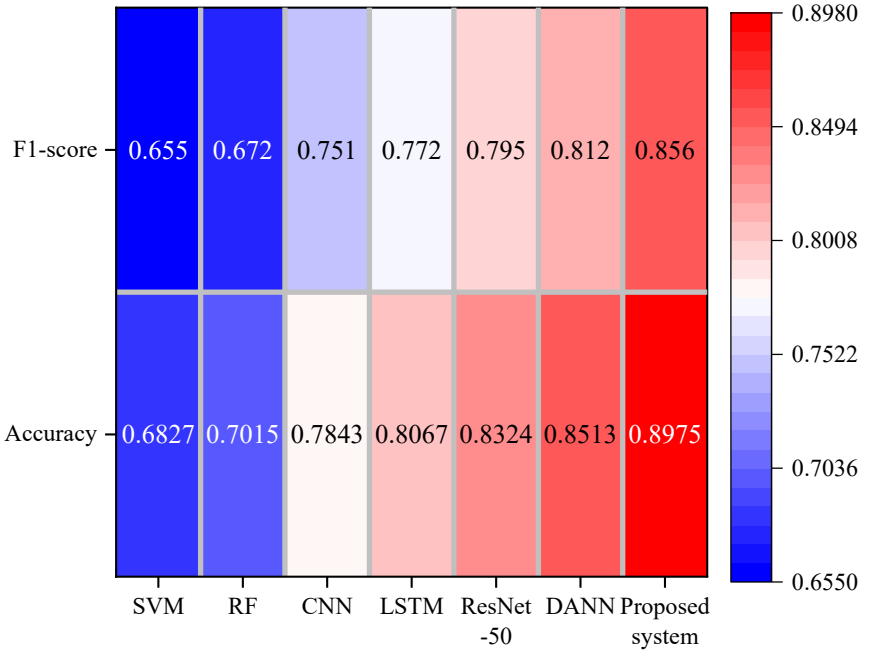


Figure 4 Performance comparison of emotion recognition models on the target domain dataset (see online version for colours)



The experimental findings from both datasets indicate that the proposed method demonstrates substantial performance benefits in emotion recognition tasks. Its recognition accuracy and stability are better than those of the comparison models in both the source domain and the target domain. The suggested system obtains an accuracy of 0.9358 and an F1 score of 0.912 on the source domain dataset, which is far better than classic ML models like SVM and RF, as well as DL models like CNN, LSTM, and ResNet-50. The suggested system has a big performance lead over the DANN model, which completely proves that TL-based feature extraction and parameter fine-tuning processes work well for recognising emotions.

In general, the performance of a model gets better as the structure gets more complex and the learning capacity increases. For example, the accuracy of recognition goes up from traditional methods like SVM and RF to deep models like CNN and LSTM, and finally to the current system that uses TL strategies. This result shows that jobs that include recognising emotions require a lot of abstraction at the feature level.

In the target domain experiments, although all models exhibited performance degradation due to domain transfer, the proposed system maintained optimal results with an accuracy of 0.8975 and an F1 score of 0.856. These figures represent improvements of 0.0651 and 0.0462 over ResNet-50 and DANN respectively. Its feature transfer and parameter fine-tuning mechanisms effectively mitigated the disparity in feature distributions between source and target domains, enabling the model to accurately recognise students' emotional states within music learning environments.

Overall, experimental results conclusively demonstrate the system's efficacy and advancement in emotion recognition. Through the introduction of the TL mechanism, the system achieves high-precision emotion recognition in the source domain while exhibiting outstanding generalisation capabilities in target domain transfer tasks. This cross-domain robustness lays a solid foundation for implementing emotion perception and feedback regulation in subsequent adaptive music teaching systems, while also providing novel insights and technical support for integrating affective computing with intelligent education technologies.

4.3 Adaptive teaching effectiveness comparison experiment

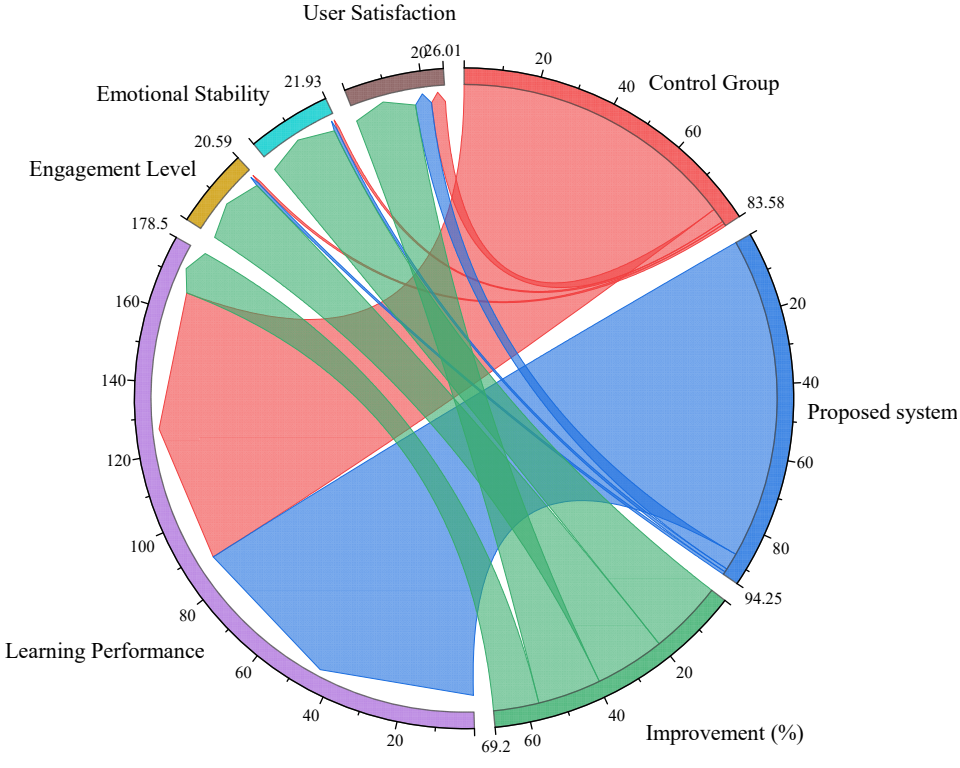
This study recruited 80 university students from three institutions as participants, comprising 41 males and 39 females with an average age of 20.4 years. All participants completed baseline assessments of musical proficiency and emotional state prior to the experiment. They were then randomly divided into two groups: an experimental group (the proposed system) and a control group (the non-adaptive system). Both groups received identical instructional content, time allocation, and task structure, with the sole distinction being the activation of an adaptive adjustment module based on emotion recognition and feature transfer within the experimental group.

The two-week experimental cycle comprised three learning tasks: melody imitation, rhythm recognition, and emotional expression. Experimental results are presented in Figure 5.

The results demonstrate that the proposed system significantly outperforms the control group in overall teaching effectiveness, with all metrics exhibiting stable and positive improvement trends. Notably, learning performance increased by 12.4%, indicating the system's substantial role in enhancing knowledge acquisition and skill consolidation. This improvement stems primarily from the system's ability to

dynamically adjust teaching pace and difficulty levels in real-time following emotional recognition. This ensures learning content aligns more closely with students' current psychological states, thereby optimising learning efficiency.

Figure 5 Overall adaptive teaching performance comparison (see online version for colours)



Engagement levels increased by 19.1%, reflecting how the system's interactive design and emotion-sensing mechanisms markedly enhanced students' focus and immersion. Utilising an emotion recognition model optimised with TL, the system automatically triggers musical feedback and visual cues upon detecting waning attention or fatigue, thereby reactivating learning motivation.

Emotional stability improved by 20.3%, a particularly noteworthy outcome. This metric directly reflects the system's capacity to maintain learners' emotional equilibrium. Compared to the control group, experimental group learners exhibited more stable emotional fluctuations during prolonged study sessions, indicating the system's distinct advantage in emotional intervention and feedback regulation.

Furthermore, user satisfaction increased by 17.4%, indicating learners' generally positive attitudes towards the system's instructional experience. Questionnaire data revealed that experimental group students perceived the system's emotional feedback as more natural, its teaching pace as better aligned with their personal states, and that the integration of visual and auditory feedback enhanced immersion.

The experimental results confirm the suggested system's multifaceted benefits: it improves learning performance, engagement, and emotional regulation. These findings

provide a solid empirical foundation for applying the system to varied learning environments.

However, to further elucidate the specific mechanisms underlying the system's functioning across different emotional states, this study also conducted a detailed analysis of the overall experimental results according to emotional dimensions. Based on the multimodal emotional data collected during the experiment, the learning process was categorised into three primary emotional states: Positive, Neutral, and Negative. Comparative results for key metrics such as learning scores, participation duration, emotional stability, and recovery time across each state are presented in Table 2.

Table 2 Teaching effectiveness across emotional states

Emotional state	*Indicator*	*Control group*	*Experimental group (proposed system)*	*Improvement (%)*
Positive (joyful/relaxed)	Learning score	86.3	90.7	+5.1
	Engagement time (min)	42.1	45.4	+7.8
	Emotional stability index	0.89	0.94	+5.6
	Recovery time (s)	4.8	4.2	–12.5
Neutral (calm/focused)	Learning score	81.6	88.9	+8.9
	Engagement time (min)	38.5	43.1	+11.9
	Emotional stability index	0.84	0.93	+10.7
	Recovery time (s)	6.7	5.2	–22.4
Negative (anxious/tired)	Learning score	73.4	84.6	+15.3
	Engagement time (min)	33.9	40.2	+18.6
	Emotional stability index	0.71	0.87	+22.5
	Recovery time (s)	9.1	6.0	–34.1

Table 2 shows the suggested system's improved instructional adaptation across emotional states. The minor improvement under positive emotions implies learners already focus and motivate in heightened emotional states, with the system mostly stabilising. Neutral learners' learning scores improve by 8.9%, demonstrating the approach improves concentration and interest.

Negative emotions cause the greatest changes: learning scores rise 15.3%, emotional stability rises 22.5%, and emotional recovery time falls 34.1%. When negative emotions are detected, the system quickly adapts instructional content and speed to help students regain psychological equilibrium and improve learning efficiency.

Experimental results show that the TL-based adaptive music teaching system can dynamically regulate multiple emotional variables. This proves its efficacy in improving learning, emotional regulation, and interaction. The system accurately detects and responds to emotional changes and optimises educational content and emotional feedback, laying the groundwork for future research into intelligent affective teaching systems.

5 Conclusions

The system is based on DNN architecture and combines feature transfer and parameter fine-tuning methods to greatly improve the accuracy of emotion recognition and the capacity to work across domains. Experimental comparisons showed that the suggested system did better than both traditional and DL models on two datasets when it came to recognising emotions. Teaching efficacy trials further confirmed the system's substantial benefits in learning performance, emotional stability, and user pleasure, notably highlighting its strong regulating and restorative effects in adverse emotional states.

Even with these good results, there are still some problems. First, the sample size for the trial was small and mostly focused on music learning areas. More research is needed to see if the results can be used in other fields or forms of learning. Second, the emotion recognition model has TL mechanism, but it still needs EEG and facial expressions to be recorded at the same time. Thirdly, the current system's adaptive strategies primarily rely on rule-based approaches and model mapping, with insufficient incorporation of reinforcement learning or generative modelling mechanisms. Consequently, there remains scope for improvement in long-term teaching interactions and the modelling of individual emotional evolution.

Future research may be pursued in three directions: firstly, expanding the sample scope to validate the system's adaptability across different disciplines; secondly, incorporating generative modelling techniques to enhance the intelligence level of emotional dynamic regulation. Third, we need to make the system easier to understand and lighter so that it can be used in real-time and for a long time in real-world teaching situations.

Declarations

All authors declare that they have no conflicts of interest.

References

Afzal, S., Khan, H.A., Piran, M.J. and Lee, J.W. (2024) 'A comprehensive survey on affective computing: challenges, trends, applications, and future directions', *IEEE Access*, Vol. 12, pp.96150–96168.

Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G. and Liu, T. (2021) 'Understanding and improving early stopping for learning with noisy labels', *Advances in Neural Information Processing Systems*, Vol. 34, pp.24392–24403.

Barakina, E.Y., Popova, A.V., Gorokhova, S.S. and Voskovskaya, A.S. (2021) 'Digital technologies and artificial intelligence technologies in education', *European Journal of Contemporary Education*, Vol. 10, No. 2, pp.285–296.

Chen, W., Liu, Y., Wang, W., Bakker, E.M., Georgiou, T., Fieguth, P., Liu, L. and Lew, M.S. (2022) 'Deep learning for instance retrieval: a survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 6, pp.7270–7292.

Cong, S. and Zhou, Y. (2023) 'A review of convolutional neural network architectures and their optimizations', *Artificial Intelligence Review*, Vol. 56, No. 3, pp. 1905–1969.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M. and Chen, W. (2023) 'Parameter-efficient fine-tuning of large-scale pre-trained language models', *Nature Machine Intelligence*, Vol. 5, No. 3, pp.220–235.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M. and Hussain, A. (2024) 'Interpreting black-box models: a review on explainable artificial intelligence', *Cognitive Computation*, Vol. 16, No. 1, pp.45–74.

Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z. and Azim, M.A. (2022) 'Transfer learning: a friendly introduction', *Journal of Big Data*, Vol. 9, No. 1, p.102.

Jagtap, S., Marne, A., Sheikh, A., Potdar, V. and Chate, P. (2023) 'Design and analysis of unlocking student emotions: enhancing e-learning with facial emotion detection', *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, Vol. 11, No. 12, pp.1824–1830.

Jiang, X., Yang, X., Huang, J., Zhou, X. and Cui, J. (2025) 'Improved deep domain adversarial neural network with joint maximum mean discrepancy for bearing multi-condition fault diagnosis', *Measurement Science and Technology*, Vol. 36, No. 3, p.036137.

Karnati, M., Seal, A., Bhattacharjee, D., Yazidi, A. and Krejcar, O. (2023) 'Understanding deep learning techniques for recognition of human emotions using facial expressions: a comprehensive survey', *IEEE Transactions on Instrumentation and Measurement*, Vol. 72, pp.1–31.

Kostakis, P. and Kargas, A. (2021) 'Big-data management: a driver for digital transformation?', *Information*, Vol. 12, No. 10, p.411.

Lara-Benítez, P., Carranza-García, M., Luna-Romera, J.M. and Riquelme, J.C. (2020) 'Temporal convolutional networks applied to energy-related time series forecasting', *Applied Sciences*, Vol. 10, No. 7, p.2322.

Li, L. and Han, Z. (2023) 'Design and innovation of audio IoT technology using music teaching intelligent mode', *Neural Computing and Applications*, Vol. 35, No. 6, pp.4383–4396.

Liao, Z., Zhang, X., He, S. and Tang, Q. (2023) 'PMP: A partition-match parallel mechanism for DNN inference acceleration in cloud-edge collaborative environments', *Journal of Network and Computer Applications*, Vol. 218, p.103720.

Min, B., Ross, H., Sulem, E., Veyseh, A.P.B., Nguyen, T.H., Sainz, O., Agirre, E., Heintz, I. and Roth, D. (2023) 'Recent advances in natural language processing via large pre-trained language models: a survey', *ACM Computing Surveys*, Vol. 56, No. 2, pp.1–40.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K. and Ray, A. (2022) 'Training language models to follow instructions with human feedback', *Advances in Neural Information Processing Systems*, Vol. 35, pp.27730–27744.

Petrolini, V. and Viola, M. (2020) 'Core affect dynamics: arousal as a modulator of valence', *Review of Philosophy and Psychology*, Vol. 11, No. 4, pp.783–801.

Seyderhelm, A.J. and Blackmore, K.L. (2023) 'How hard is it really? Assessing game-task difficulty through real-time measures of performance and cognitive load', *Simulation & Gaming*, Vol. 54, No. 3, pp.294–321.

Takale, D.G., Mahalle, P.N. and Sule, B. (2024) 'Advancements and applications of generative artificial intelligence', *Journal of Information Technology and Sciences*, Vol. 10, No. 1, pp.20–27.

Thaher, T., Zaguia, A., Al Azwari, S., Mafarja, M., Chantar, H., Abuhamdah, A., Turabieh, H., Mirjalili, S. and Sheta, A. (2021) 'An enhanced evolutionary student performance prediction model using whale optimization algorithm boosted with sine-cosine mechanism', *Applied Sciences*, Vol. 11, No. 21, p.10237.

Wang, P-T., Sheu, J-S. and Shen, C-F. (2023) 'Real-time hand movement trajectory tracking with deep learning', *Sensors and Materials*, Vol. 35, No. 12, pp.4117–4129.

Yan, J., Du, C., Li, B., Zhou, X. and Liu, Y. (2025) 'Cross-database facial expression recognition based on multi-feature representation multi-layer domain adaptive fusion network', *Engineering Applications of Artificial Intelligence*, Vol. 155, p.110995.

Zhang, L. and Gao, X. (2022) 'Transfer adaptation learning: a decade survey', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 35, No. 1, pp.23–44.

Zhu, X., Huang, Y., Wang, X. and Wang, R. (2024) 'Emotion recognition based on brain-like multimodal hierarchical perception', *Multimedia Tools and Applications*, Vol. 83, No. 18, pp.56039–56057.