Asia Pacific
Academy of Science Pte. Ltd.

ORIGINAL RESEARCH ARTICLE

Application and algorithm optimization of music emotion recognition in piano performance evaluation

Yao Zhang[1], Delin Cai[2], Dongmei Zhang[3,*]

Academy of Arts, Yunnan Minzu University, Kunming 650031, Yunnan Province, China

*** Corresponding author:** Dongmei Zhang, 041101@ymu.edu.cn

ABSTRACT

In the current research, we integrate distinct learning modalities—Curriculum Learning (CL) and Reinforcement Learning (RL)—in an attempt to develop and optimize Music Emotion Recognition (MER) in piano performance. Classical approaches have never been successful when applied in the field of determining the degree of emotion in the music of the piano, owing to the substantial complexity required. Addressing this particular issue is the primary motivation for the present endeavour. In an approach that's comparable to how human beings acquire information, it trains the RL agent CL in phases; such an approach improves the student's learning model in understanding the diverse emotions expressed by musical compositions. A higher rating of performance can be achieved after learning the model to recognize emotions more effectively and precisely. A set of piano melodies with emotional content notes has been included in the EMOPIA repository for use when conducting the process of evaluation. In order to benchmark the proposed approach with different models, parameters including R^2, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) were deployed. Studies indicate that the recommended approach accurately recognizes the emotions expressed by piano-playing music. In challenging tasks like MER, the significance of implementing the CL paradigm with the RL has been emphasized using the outcomes mentioned earlier.

Keywords: Curriculum Learning; Music Emotion Recognition; piano music; Machine Learning; Reinforcement Learning; MBE; RMSE

1. Introduction

The piano player practices evaluation-based Music Emotion Recognition (MER) is a challenging MAT problem to address due to the fact that it requires both proficiency in technology and emotional empathy. Creativity and expressive emotion on the part of the musician are essential throughout each pianist's play[1]. The method of evaluation additionally becomes more challenging by integrating all of these variables. Instead of simply paying attention exclusively to emotion recognition, a MER model ought to know how these feelings affect the whole experience along with performance level. Piano plays are highly personal in their expression of emotion; minor variations in rhythm, behavior, and clarity can have significant effects on the standards used to evaluate how music expresses itself. The intricate and distinct nature of expressive emotion shows the most effective task for MER in playing music performances. There is an inherently complex connection between musical properties and the emotions they express, and standard MER techniques have always had problems

ARTICLE INFO
Received: 12 December 2023 | Accepted: 19 January 2023 | Available online: 24 January 2024

CITATION
Zhang Y, Cai D, Zhang D. Application and algorithm optimization of music emotion recognition in piano performance evaluation. *Environment and Social Psychology* 2024; 9(4): 2344. doi: 10.54517/esp.v9i4.2344

accurately expressing this. Based on the musician, even an identical recording of piano music may cause entirely distinct emotions, which those methods are unable to express[2].

A feasible answer to the issues resulting from MER is Machine Learning (ML), which relies on a machine's capacity to learn from historical information in order to make accurate forecasts. Algorithms used for machine learning may be trained on how to recognize musical indications and rhythms related to particular emotions by developing models on an extensive collection of compositions by musicians[3]. Within the scope of ML, Reinforcement Learning (RL) is commonly recognized for its excellent results on complex tasks obtained by learning via ongoing interactions with the natural world. Real-life prototypes learn constantly in order to enhance their listening skills, which gives them an innate capacity to identify complex emotional trends in musical instruments[4]. With all of the data collected and the unique behavior of musical emotions, these algorithms continue to have trouble with the MER objective. They are not effective because they demand a great deal of initial training data and can't apply to different types of musical instruments, so it's essential to maximize the RL models. Hence, they are capable of managing sophisticated MER applications. The primary objective of this refinement is to improve RL's learning system more reliably so that it can more effectively deal with the varied and complex nature of musical emotions. Learning the RL method sequentially with the convoluted method of musical emotions could be more effective, resulting in a more comprehensive and infinite learning method.

In order to develop the RL to enhance MER in playing the piano rehearsals, the present investigation recommends employing Curriculum Learning (CL)[5]. This makes it feasible for the successful implementation of the RL approach in the MER challenge[6]. By gradually boosting the degree of detail of the functions that were given to the RL mathematical models, the Curriculum Learning Optimised Reinforcement Learning (CL-RL) method assists the RL agents in comprehending more about their current level of complexity. Researchers analyze the proposed hypothesis on the EMOPIA dataset, which includes numerous types of piano lesson plays annotated with various levels of emotion. The outcome assessment of the proposed model and the other models was done using a system of measurement such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2), and the findings proved the model's size in precisely identifying emotions in piano music. The results of this study confirm the theoretical potential of CL and use it to improve the performance of RL methods in the application field of MER, as the CL-RL model outperformed standard CL, RL, and RNN algorithms.

2. Literature review

The significance of assessing timbre in vocal music performances together with its emotional expression is highlighted[6–10]. By receiving data on audio, their research team implemented Support Vector Machine (SVM) methods to reduce it. This illustrates an appropriate approach for recognizing the sounds of voices in compositions for music. The results they reported are in contrast to what is typically understood about the significance emotions have on performances by musicians. Research the performances of music thoroughly by employing techniques for tracking feedback from the audience[7,11–13]. The method they use uses a multidimensional paradigm that precisely recognizes the emotional elements of audio recordings by using feelings of arousal and valence as the most important signals. Differences, as well as similarities in the emotional effect of identical music, have been determined by measuring these features throughout various performances as well. Early learners can explore not one but two audio-centric techniques[14–18]. Researchers designed a time-dependent and partitioned system by features to be extracted using Convolutional Neural Networks (CNNs) and automatically manipulating the progression of time. Furthermore, researchers invented a complete framework that applies CNN and attention optimization approaches. By introducing the methods mentioned earlier to the Yingcai Piano Performance Evaluation Phase III Dataset, researchers are able to demonstrate that accurate musical instruments can be manufactured in order to evaluate piano playing performances regardless of whether audio inputs are unavailable.

In order to better understand the psychological elements in musical instruments, investigators invented an innovative approach to categorize instruments used for music[19–23]. Researchers extract features, including MFCC and Chroma STFT, from libraries of music that are instrumental using Deep Learning (DL) approaches. Many of these features are categorized further based on numerous emotional factors. In particular, their study demonstrates that deep RNNs outperform conventional ML algorithms when identifying emotions in music based on instrument categories, which is a key finding in the relevant literature.The need to give piano students a way to evaluate their playing performance is highlighted, so they can get helpful feedback[24–27]. The real-time recognition of single notes and the non-real-time recognition of multiple notes are both handled by this method. It uses techniques like local energy endpoint detection, which enhances the system's real-time operation and capacity to handle different scenarios successfully (**Table 1**).

Table 1. MER context-based features and music content.

Reference	Type	Features	Methodology	Accuracy
Setiawan et al.[28]	AMG	MARSYAS	SVM	56.18%
Irrinki[29]	Sound Track	CHROMD, SPITCH	Decision Tree	70%
Lakshmi et al.[30]	CAL500	PERCTO	TEML + SVM	56.4%
Madhavi and Lavanya[31]	Chinese & Western	RHYT	BAYN	74.9%
Nanduri et al.[32]	Mood Swings Lite	OBSC	LDS KALFM	2.88%
Srinivas et al.[33]	Western Pop	RPEAKVAL	OAO FSVM	37%
Sugumar et al.[34]	Clips	OBSC	SVM	67.5%
Kuchibhotla et al.[35]	TV Theme Tunes	TEMP	SVM	80.99
Chintalapudi et al.[36]	Piano and Vocal	34 MFCCs	DTM	60%
Fernandes Dimlo et al.[37]	Movie Soundtrack	HARM	k-NN	59.4%
Nallapu et al.[38]	Chinese Songs	TF-IDF	SVM-SMO	61.5%

3. Background

3.1. Curriculum Learning (CL)

The idea behind CL was a set of training intervals represented by the symbol F[1]. During those intervals, each $R_k(y)$ phase modifies the underlying data distribution $G_k(y)$. A dataset variable, $D(y)$, is usually represented by a pair (x,y) in supervised learning scenarios (Equation (1)):

$$\begin{gathered} F = \langle R_1, \ldots, R_2, \ldots, R_N \rangle \\ R_k(y) \propto G_k(y) D_k(y), \forall y \in S \end{gathered} \tag{1}$$

which must satisfy:

(a) A systematic expansion of the type and depth of information contained in the training subset.

(b) The correlation between the two variables indicates that the training samples are progressively added (Equation (2)):

$$G_k(y) \leq G_{k+1}(y) \tag{2}$$

(c) Adjusting the weights of each sample so they are consistent which prepares the dataset for training (Equation (3)):

$$R_N(y) = D(y) \tag{3}$$

The development of CL has been so remarkable that it is now being used at levels above its initial standards, including data, tasks, and models. This development has led to the emergence of a three-stage framework: assessing the model's difficulty, developing a training approach, and evaluating the model. Integrating these steps into a unified framework ensures a thorough and flexible application of CL to different ML applications by addressing the critical traits of task, data, and model.

Let A be the primary dataset, and U be the task set. The training subset is represented by $a,$ and the training process comprises a difficulty assessor (D_a), a training organizer (O), a performance metric (P_m,), and the model (M_a). In the task set U, a subtask refers explicitly to a particular task.

Phase 1: Assessment of difficulty: During this stage, standards for evaluating the complexity of samples in A are defined. To determine how challenging the samples are, a complexity evaluator called D_a is employed. A training sequence R is constructed using this evaluation, with examples adapting in difficulty from y_{simple} (the easiest) to $y_{complex}$ (the most challenging) (Equation (4)):

$$R = \{y_{simple}, \dots, y_j, \dots, y_{complex}\} \text{ for } j < m', y \in A \tag{4}$$

Phase 2: Formulation of training schedule: The training organizer O is used to establish the model's training progression criteria. In the t^{th} training cycle, two subsets are formed from the sequence R: one that is task-oriented (a_{job}) and another that is data-focused (a_{info}) (Equations (5) and (6)):

$$a_{info}^t = \{y_1, y_2, \dots, y_k\}, k \le p \tag{5}$$

$$a_{job}^t = \{subtask_1, subtask_2, \dots, subtask_k\}, k \le p \tag{6}$$

Phase 3: Evaluation of the model: Following the creation of the training subset a_{info} in the previous phase, the current state and learning progress of the model are assessed using the performance metric P_m. The acquired knowledge from this evaluation is subsequently communicated to both the training organizer O and the difficulty assessor D_a. The feedback provided is used to periodically reassess the complexity of the samples and the training subset a_{info} is updated accordingly. This phase is crucial for the CL of the network's model, demonstrating a gradual evolution from the initial model configuration $M_a 1$ and progressively adapting the parameters and structure of the network until the final, comprehensive model M_a is employed for training $\langle M_a 1, \dots, M_a t, \dots, M_a \rangle$. **Figure 1** depicts the CL model.

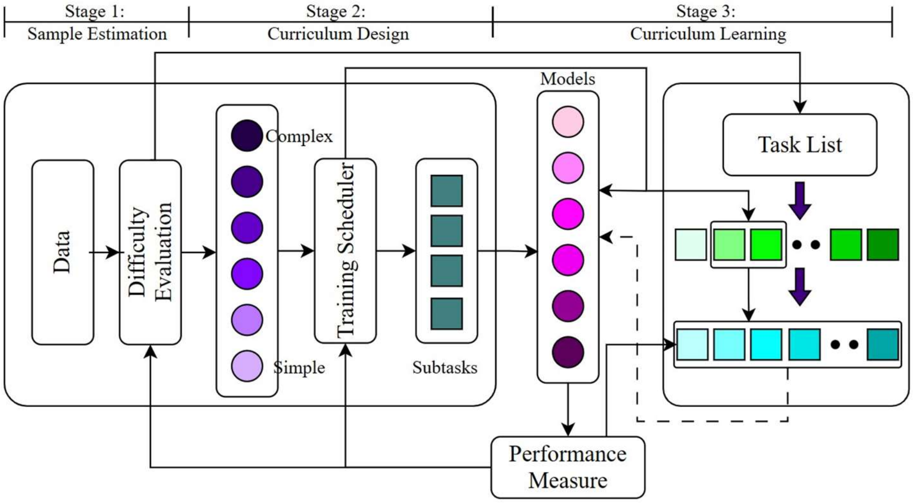

Figure 1. The CL framework.

Algorithm 1 presents the CL Framework, and it starts with the difficulty assessment of dataset A using the evaluator D_a, which produces a difficulty-ranked list R. The list R is then employed by the organizer O to assemble the initial training set a_1. During each iteration, numbered 1 to k, the model's performance, denoted M, is evaluated using the metric P_m. Subsequently, D_a reassesses the sample difficulty considering performance, updating the list to R'. Simultaneously, O uses R' to construct a new training set a_2, aligning it with the model's performance and the revised difficulty rankings. The training process is repeated with the updated training set until the model achieves convergence.

```
Algorithm 1: The Curriculum Learning
Input:
  • Dataset $A = \{(x_j, y_j)\}_{j=1,2,\dots,m}$
  • Initial training set $a$
  • Difficulty Assessor $D_a$
  • Ranked List $R$
  • Training Organizer $O$
  • Performance Evaluator $P_e$
  • Model $N$
Output:
  • Optimized Model $N^*$
Procedure:
Step 1.   Generate a difficulty-ranked list $R$ from $A$ using $D_a$ $R\{\text{simple}, \dots, \text{moderate}, \dots, \text{complex}\} = D_a(A)$.
Step 2.   Formulate the initial training subset $a$ using $O$ based on $R$ and $D_a$: $a = O(R, D_a)$.
Step 3.   For each iteration $s$ from 1 to $k$ :
Step 4.   Update the training subset $a$ using $O$ based on the updated list $R$ and performance $P_e$: $a = O(R, P_e)$.
Step 5.   Proceed with training until convergence is not achieved for $P_e$ Epochs: Train $N$ using $a$ and $O$.
Step 6.   Reassess the difficulty list $R$ using $D_a$ based on $A$ and $N$: $R = D_a(A, N)$.
Step 7.   Evaluate the current model performance $P_e$ with $P_e(A, N)$.
Step 8.   Continue until the model $N$ converges, resulting in the optimized model $N^*$.
```

3.2. RL environment for MER

The RL environment for recognizing emotions in piano performances is configured around four key elements: State, Action, Reward, and Policy Function. These components are integral in training the RL agent to interpret and categorize emotions accurately.

A. State space in MER: The state generation process in this model is designed to capture both the current features and the historical and emotional context of piano performances. It is achieved through a sliding window mechanism, as illustrated in the proposed model's structure diagram. At any given time t, the state, denoted as $\mathbb{S}_t$, is comprised of the current feature set F_t and the aggregated emotional states from the preceding time steps, represented as $\mathcal{E}_{t-1}$. The state is thus formulated as Equation (7):

$$\mathbb{S}_t = [\mathcal{E}_{t-1}, F_t] \tag{7}$$

Here, F_t is the current features, including melody, harmony, and rhythm. The aggregated emotional states $\mathcal{E}_{t-1}$ are measured through a max-pooling operation over the emotional states from a window of past sequences, expressed as Equation (8):

$$\mathcal{E}_{t-1} = \text{MaxPool}\ ([E_{t-W}, \dots, E_{t-1}]) \tag{8}$$

where, E_{t-1} represents the emotional state output at time $t-1$, and W is the window size. For the initial state $\mathbb{S}_1$, where no past emotional context is available, the aggregated emotional state $\mathcal{E}_0$ is initialized randomly, and F_1 represents the first observed feature set (Equation (9)):

$$\mathbb{S}_1 = [\mathcal{E}_0, F_1] \tag{9}$$

The state set $\{\mathbb{S}_t\}$, a comprehensive view of the music's present and past emotional landscape facilitates emotion recognition.

B. Actions in MER: In this research of the RL model for MER in piano performances, the set of actions is specifically designed to classify the current music segment into one of several emotion groups. These categories encompass a range of emotions typically conveyed in music, such as "Joy", "Sadness", "Anger", "Fear", "Surprise", "Disgust", and "Neutral". At each time step $\boldsymbol{i}$, given the current state $\boldsymbol{State_i}$, which combines the aggregated emotional state $\boldsymbol{\mathcal{E}_{i-1}}$ and the current feature set $\boldsymbol{F_i}$, the RL agent selects an action $\mathbb{a}_i$. This selection is guided by the agent's policy function $\boldsymbol{\pi}(\mathbb{a}_i \mid \mathbb{S}_i)$, which determines the most suitable emotional type based on historical and emotional context and the present musical features. Once an action $\mathbb{a}_i$ is chosen, it triggers the computation of a reward. This reward reflects the fitness of the selected emotion class about the current state and provides feedback from the environment to the agent. This feedback is integral for learning and updating the parameters within the Deep-Q network.

C. Reward function and Policy Learning in MER: In this proposed model, unlike standard active learning methods that rely on ambiguity measurements or information density, this work employs data uncertainty as the guiding principle for the RL policy in the context of MER. This uncertainty, indicative of the difficulty in accurately categorizing a musical segment's emotion, is quantified using a loss metric.

D. Reward function: The reward function in this model is formulated to reflect the accuracy of emotion recognition. It is computed using the cross-entropy loss between the predicted and actual emotion labels. The reward at time step i, denoted as $\mathbb{R}_i$, is given by Equation (10):

$$\mathbb{R}_i = 1 - \sum_{j=1}^{N} y_{ij} \log\left(\mathbb{a}_{ij}\right) \tag{10}$$

Here, N represents the number of emotion types, $\mathbb{a}_{ij}$ is the probability of the selected action for emotion type j at time step i, and y_{ij} is the valid emotion label for set j at time step i.

E. Future reward calculation: Considering the sequential nature of music, the future rewards are discounted over time to calculate the expected cumulative reward, Q. This is expressed in Equation (11).

$$Q^* = \max_{\pi} \mathbb{E}[\mathbb{R}_i + \lambda \mathbb{R}_{i+1} + \lambda^2 \mathbb{R}_{i+2} + \cdots \mid \pi, \mathbb{S}_i, \mathbb{a}_i] \tag{11}$$

In this Equation (11), λ is the discount factor, indicating the importance of future rewards, and π is the policy function.

F. Policy learning: The RL agent aims to progressively learn the patterns of emotional changes in piano music by maximizing the cumulative reward received during interaction with the environment. The policy

$\pi(\mathbb{S})$ determines the action in a given state $\mathbb{S}$. Since there is no likelihood of state transitions and the reward function is data-dependent, this frames this as an RL problem and uses Q-learning to solve it.

4. Proposed methodology

4.1. Need for integrating CL with RL for MER

RL models often resist the tasks of MER due to the complexity and collection of musical emotions. Some of these challenges include lacking the ability to differentiate between subtle shifts in emotion, having to spend more time in training due to the comprehensive collection of musical expressions, and not being able to apply what you learn to other styles of music. The RL model tries to understand and correctly identify emotions due to the complexity of these components. By incorporating CL into the RL framework, these concerns can be effectively handled by streamlining the learning process. CL allows for gradually initiating progressively intricate musical elements, evolving from more straightforward to complex emotional emotions. In addition to improving the RL model's learning performance, this approach lays a solid groundwork for standardization. Through CL, the RL model learns more about emotional shifts over time, improving its accuracy and versatility when recognizing emotions in distinct types of piano music.

A. Music emotion recognition model

Two directions can be roughly split into available MER study articles. The two investigated directions are Song-Level MER (SLMER) and Music Emotion Variation Detection (MEVD). The abbreviation "SLEMR" describes the technique of determining an aggregate emotion tag for one particular song. It must be done that researchers consider the ongoing and unpredictable method of emotion when researching investigations into MEVD, even though MEVD regards the emotion of music as an ever-evolving method.

The results presented in **Table 2** highlight MER's most commonly used MER models. Emotion models associated with music have been identified by "Music" in the "Application Domain" section, while "General" represents emotion models appropriate to the overall historical context. The adaptable nature of computational emotion models presents them as a good option for multi-modal MER sentiment analysis. Whenever it has to do with precisely describing the emotions triggered by music, absolutely nothing matches a musician's specific emotion model. While relating to the "Emotion Conceptualization" section, the term "Categorical" mentions the categorical emotion model; however, the expression "Dimensional" symbolizes the dimensional version of the MER model (**Figure 2**). A growing number of researchers have presented queries regarding the accuracy of the categorical emotion model, which has caused the adoption of dimensional emotion models in recent years. In contrast to "Induced", which defines an emotion systematically developed, "Perceived" expresses an emotion that has been noticed. In general, music metadata must be present to understand an emotional response from an example of music.

Specifying the field of study, as depicted in **Figure 2**, needs to be done to select emotion models and datasets to define the study focus. The following sections provide a quick summary of the subject in general, along with additional information and standard emotion models and datasets.

Table 2. The plan of action for study for MER.

Methodology	Conceptualizations of emotion	Synopsis
Song-level MER	Categorical approach	Determine music emotional categories.
	Dimensional approach	Identify musical emotion frequencies.
MEVD	Categorical approach	Find music's dynamic absolute emotion variation.
	Dimensional approach	Music's dynamic dimensional emotion variation prediction

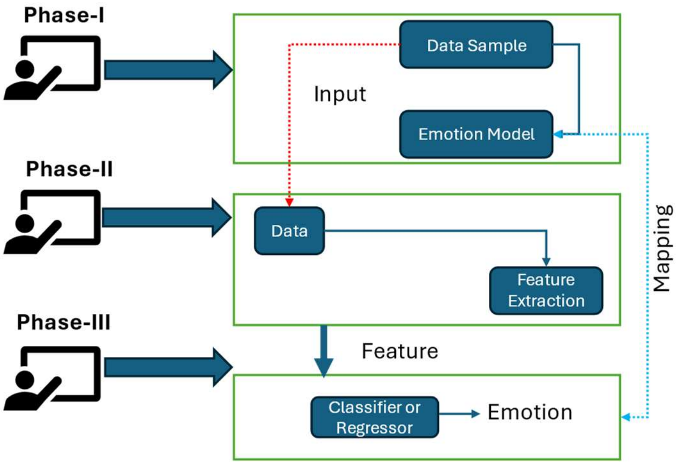

Figure 2. Model of music emotion recognition.

4.2. Dataset description

This study applies the EMOPIA dataset, which contains 1087 clips representing 387 separate piano solo performances (**Table 3**). The collection comprises soundtracks from films, popular songs from Korean and the Western world, Japanese animation, and original musical compositions. To maintain emotional and melodic coherence, each clip in the dataset is manually divided at cadential arrivals. Clarity of emotional expression and recording quality are the primary considerations in this selection procedure. Using a four-class taxonomy based on valence and arousal levels—HVHA, HVLA, LVHA, and LVLA—the emotion annotations in EMOPIA are categorized according to Russell's Circumplex model of affect.

Table 3. Dataset description.

Feature	Details
Total clips	1087 clips from 387 piano solo performances
Genres	Japanese anime, Korean, and Western pop songs cover movie soundtracks.
Emotion model	Russell's Circumplex model: HVHA, HVLA, LVHA, LVLA
Annotation	Conducted by the first four authors
Transcription	A high-resolution piano transcription model used
Encoding formats	MIDI-like, REMI, CP
Accessibility	Metadata, annotations, and MIDI available; audio via YouTube links

4.3. Feature extraction model using mel-frequency cepstral coefficients (MFCC)

Incorporating MFCCs into the learning model for MER extracts and analyzes the acoustic features of piano performances in the dataset. Convert the time-domain signal of each audio clip into the frequency domain (Equation (12)):

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-\frac{i2\pi}{N}kn} \tag{12}$$

where $X(k)$ is the fourier transform, $x(n)$ is the time-domain signal, N is the number of points in the Fourier transform, and k is the frequency index. Apply a set of filters, known as the Mel filter bank, to the power spectrum to extract the spectral energy. The Mel scale is used to approximate human auditory perception (Equation (13)):

$$M(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right) \tag{13}$$

where $M(f)$ is the Mel-scaled frequency and f is the linear frequency. Convert the log Mel-scaled spectrum into time coefficients, yielding MFCC (Equation (14)):

$$C(m) = \sum_{k=1}^{K} \text{Log } S(k) \cdot \text{COS}\left[\frac{\pi m}{K}(k - 0.5)\right] \tag{14}$$

where $C(m)$ is the m-th MFCC, $S(k)$ is the Mel-scaled spectral components, and K is the total number of Mel-scale filters. Next, select the feature vector 'F' using the first 12 coefficients, which are considered to be the most relevant for capturing the audio signal's characteristics (Equation (15)):

$$F = [C(1), C(2), \dots, C(12)] \tag{15}$$

Further, the Delta (ΔC) and Delta-Delta coefficients ($\Delta^2 C$) are used to capture the rate of change of MFCCs over time. The Delta coefficient's information provides the rate of change in MFCCs over time and is computed as the first derivative of the MFCC using Equation (16):

$$\Delta C(m) = C(m + 1) - C(m - 1) \tag{16}$$

Then, the Delta-Delta coefficients or the second-order derivatives of the MFCC are computed using Equation (17):

$$\Delta^2 C(m) = \Delta C(m + 1) - \Delta C(m - 1) \tag{17}$$

The complete feature vector for the dataset's audio clip is represented as a concatenation of the selected MFCC, their Delta, and Delta-Delta coefficients. The Complete Feature Vector 'V' is then described as Equation (18):

$$V = [F, \mathit{\Delta} F, \mathit{\Delta}^2 F] \tag{18}$$

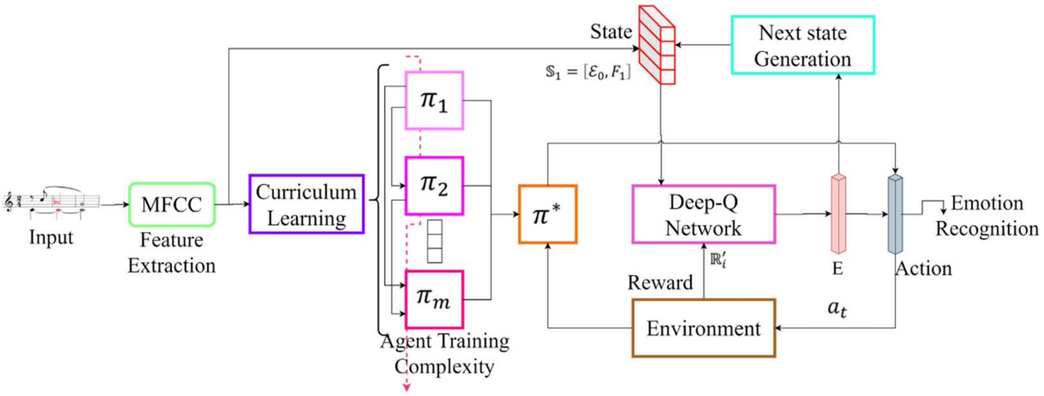

Figure 3. Proposed CL-RL model.

4.4. CL-integrated RL network model for emotion recognition

In the proposed model agent (**Figure 3**), the environment for the RL consists of segmented piano performances. Based on MFCC, which captures the temporal auditory characteristics necessary for emotion categorization, each segment offers a state. At each timestep t, the state S_t integrates the current feature set F_t with the historical context of emotional states E_{t-1}.

For the curriculum aspect of the model, this study introduces a modified metric to gauge the complexity of musical segments. The metric calculates segment difficulty based on emotional transition frequency (Equation 19).

$$d_{mc}(m_i) = \frac{N_{es}(m_i) + N_{ipp}(m_i)}{N_{\mathrm{seg}}(m_i) + N_{vp}(m_i)} \tag{19}$$

where $N_{es}(m_i)$ represents the number of emotion shifts, $N_{seg}(m_i)$ denotes the total number of segments and $N_{vp}(m_i)$ is the difference penalty that smooths the difficulty of scoring. This curriculum is operationalized through a training scheduler that organizes the dataset into stratified buckets based on the difficulty scores $\{D_1, \cdots, D_T\}$. The agent begins training on the simplest bucket D_1, and upon reaching a pre-defined performance threshold or after a set number of epochs, the agent is progressively exposed to more complex buckets.

To align this curriculum-based approach with the RL framework, this paper considers the integration's impact on the traditional RL equations. In the curriculum-integrated RL model, the reward function $\mathbb{R}_i'$ is modified to include a difficulty factor $D(m_i)$, reflecting both the accuracy of emotion recognition and the complexity of the musical segment (Equation (20)):

$$\mathbb{R}_i' = D(m_i) \cdot \left(1 - \sum_{j=1}^{N} y_{ij} \log\left(\mathbb{a}_{ij}\right) \right) \tag{20}$$

The factor $D(m_i)$ adjusts the reward according to the complexity of the segment, assigning greater values to more complex segments. The design aims to incentivize the learning model to handle difficult sections better, improving overall learning efficiency.

The policy functions $\pi_1, \pi_2, \ldots, \pi_k$ encapsulate the incremental complexity stages defined by the curriculum. The policy π_{mm} is a meta-model that is iteratively updated, synthesizing the distilled knowledge from more straightforward policies to handle the present complexity level. As the complexity of tasks increases, the Q-learning update rule, which traditionally is defined as Equation (21), is adapted to account for the curriculum. In the above equation, α_t and γ_t are the learning rate and discount factor at curriculum level t, respectively. These parameters are tuned during the progression through different complexity levels in the curriculum, enabling a customized learning process that adapts to the increasing difficulty of tasks.

$$Q(S_t, a_t) \leftarrow Q(S_t, a_t) + \alpha_t \left[\mathbb{R}_{t+1} + \gamma_t \max_a Q(S_{t+1}, a) - Q(S_t, a_t)\right] \tag{21}$$

The implementation of this model aims to produce an RL agent that demonstrates improved recognition of emotional expressions in music. The model is designed to be robust and capable of adjusting to diverse musical contexts and complexities, thus offering a promising MER research path. This CL-enhanced RL approach is expected to facilitate the development of a more effective and versatile model for interpreting emotional cues in piano music. The following algorithm presents the above model in more explanatory steps:

Algorithm 2: CL-Enhanced RL for MER

Input:

- D: Dataset of piano performance segments from the EMOPIA dataset
- W: Window size for historical, emotional context
- T: Number of curriculum buckets
- E: Epoch threshold for progression

- α, γ: Initial learning rate and discount factor
- λ: Discount factor for future rewards

Output:

- $\mathcal{C}$: Emotion classifications for each segment in the dataset

Procedure:

Step 1: Initialize

- Extract features F_t from D using MFCCs
- Define initial state $S_1 = [E_0, F_1]$ with E_0 randomly initialized
- Pre-calculate difficulty scores $d_{mc}(m_i)$ for all segments in D
- Initialize policy function π

Step 2: Bucket Formation

- Sort segments into buckets $\{D_1, D_2, \ldots, D_T\}$ based on $d_{mc}(m_i)$
- Start with the first bucket D_1 as the current training set

Step 3: Training Loop

- For each epoch e until the epoch threshold E is reached:
- For each segment m_i in the current training set:
- Generate state S_t using $[E_{t-1}, F_t]$
- Select action a_t according to current policy $\pi(a_t \mid S_t)$, representing the emotion classification
- Calculate reward $\mathbb{R}_t$ using the reward function
- Update Q-value using the adapted Q-learning rule
- If the performance threshold is met or after E epochs:
- Merge the next bucket D_{next} into the current training set
- Adjust α and γ based on the new difficulty level

Step 4: Policy Update

- Update policy π using the Q-values and expected cumulative reward
- If the model converges or the final bucket is reached, finalize the policy π^*

Step 5: Classification and Termination

- Apply the finalized policy π^* to classify emotions for each segment in D, producing the output $\mathcal{C}$
- Evaluate the classification accuracy on a validation set
- If the validation performance is satisfactory, end training
- Else, refine the training set by recalculating difficulty scores and repeat step 3

5. Experimental setup

The experiments were performed in a system configured with an Intel Xeon E5-2698 v4 processor with 20 cores and 128GB DDR4 RAM, supported by an NVIDIA Tesla V100 GPU with 32 GB memory and a 2TB SSD for storage. Software configuration encompassed Ubuntu 20.04 LTS and Python 3.8, utilizing TensorFlow 2.4 and PyTorch 1.7 as ML frameworks. For evaluation, regression algorithms underwent 10-fold cross-validation. Performance metrics involved the R^2 and MAE. Arousal and valence annotations in the dataset were scaled between [–0.5, 0.5] before model training. The model was compared against standalone CL, RL, and Recurrent Neural Network (RNN) models and the proposed CL-RL integrated model to ascertain its efficiency in MER.

For the assessment of the model's performance on the EMOPIA dataset, which is categorized into four quadrants corresponding to the classes HVHA, HVLA, LVHA, and LVLA, a 10-fold cross-validation method was employed. At the start of learning the ML algorithm, the data set's emotional and physiological variables were transformed into values between –0.5 and 0.5. For the purpose of measuring the framework, authors use the following system of measurement: R^2, MAE, RMSE, and accuracy (Acc) are the performance metrics that are performed to rate the predictive ability of the model. The most appropriate systems for comparing the recommended model are the CL, RL, and RNN models, which function independently of one another, as well as a proposed CL-RL unified model that acts as an accepted standard for MER capabilities. **Table 4** gives an explanation of the values of parameters that are required to train the models that have been investigated.

Table 4. List of parameter used for CL models.

Parameter	CL	RL	RNN	CL-RL integrated
Learning rate	0.01	0.005	0.01	0.007
Batch size	32	64	32	48
Number of epochs	100	150	120	150
Optimizer	Adam	RMSprop	Adam	Adam
Loss function	MSE	Cross-entropy	MSE	Cross-entropy
Number of layers	3	4	2	4
Hidden units per layer	128	256	64	200
Activation function	ReLU	ReLU	Tanh	ReLU
Dropout rate	0.3	0.2	0.5	0.25
Regularization technique	L2	L1	L2	L2

Analysis

Figure 4 demonstrates the findings of the research, which illustrate that the CL-RL unified model that was recommended had enhanced the accuracy of the MER. This boost is more visible when compared to the intricate emotional features that have been categorized by each of the 4 quadrants of the EMOPIA dataset.

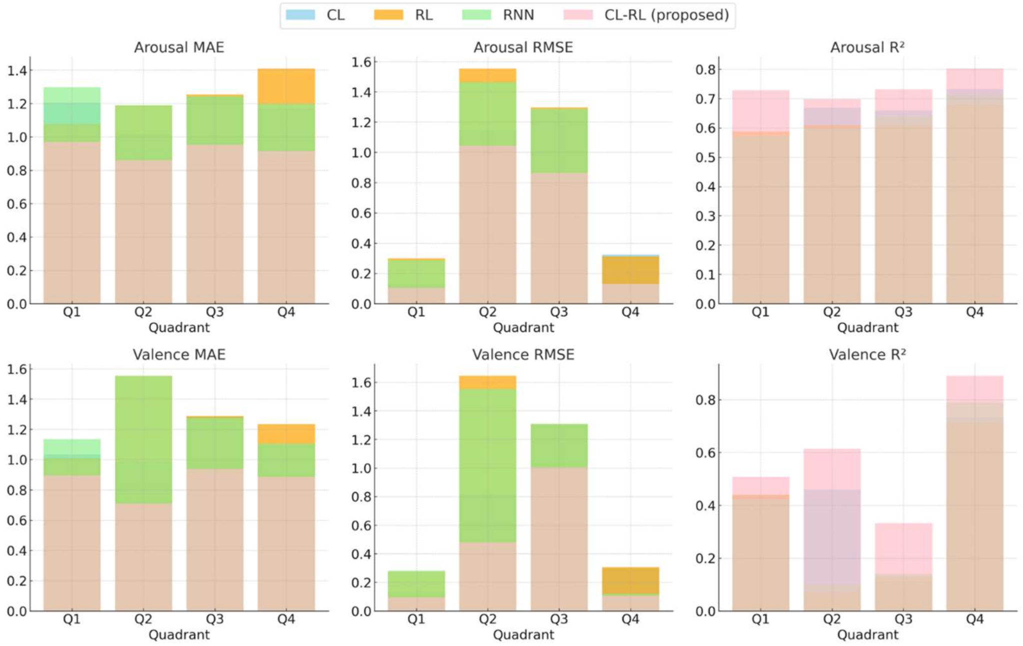

Figure 4. Performance analysis of EMOPIA dataset.

These findings illustrate that concerning each approach checked, the CL-RL model outperformed the standalone CL, RL, and RNN models on the basis of R^2, MAE, and RMSE. The model that was recommended had the highest average R-squared (R^2) values for the levels of arousal and valence in Quadrant 1 (Q1), which corresponds to High Valence and High Arousal (HVHA), while also having the lowest values for MAE and RMSE. It implies it is capable of forecasting the state of emotions with outstanding precision and developing an emotional bond with humans.

This higher efficiency has been recognized in the RL framework's CL score combination. CL ensures that the RL agents don't face the whole challenge of the problem all at once by adding musical complexity to the learning process in stages. For example, the model has shown excellent improvement in performance, as shown by R^2 measure, compared to other models in Q4 (LVLA-low valence, low arousal). This indicates that the model can consistently classify the nuanced emotions that are typically subtle and difficult to discern. This improvement is authorized to the gradual learning approach.

Furthermore, the CL-RL model's decreasing MAE and RMSE values performance in all quadrants indicates a minimal discrepancy between the projected emotional states and the actual annotations. The reduced error margins indicate that the CL approach has enhanced the model's predictive capabilities.

By adhering to a directed and regulated learning trajectory, this approach improved the agent's feature extraction and interpretation capabilities. The increase in R^2 Values from Q1 to Q4 for the CL-RL model were opposite to the other models, indicating that combining CL with RL had, in turn, enhanced the RL agents' ability to comprehend and generalize information in a more sophisticated manner.

The CL-RL model outperformed the other two models in terms of accuracy when tested across all four quadrants of the EMOPIA dataset **(Figure 5**). Quadrant 1 (Q1) accuracy for this proposed model is 93.51%, higher than that of the CL, RL, and RNN models. With an accuracy of 97.12% in Quadrant 2 (Q2), the CL-RL model proved outstanding improvement, as Q2 is associated with complex and less powerful feelings. This performance is most welcoming. Quadrant 3 (Q3) depicts emotions with intense arousal and low valence, and the CLRL model has an accuracy of 95.36% in this Quadrant. Emotions in the fourth quadrant (Q4) are mild and often hard to detect because of their low valence and arousal levels; in this quadrant, the CL-RL model achieves a recognition accuracy of 93.24% for these emotions.

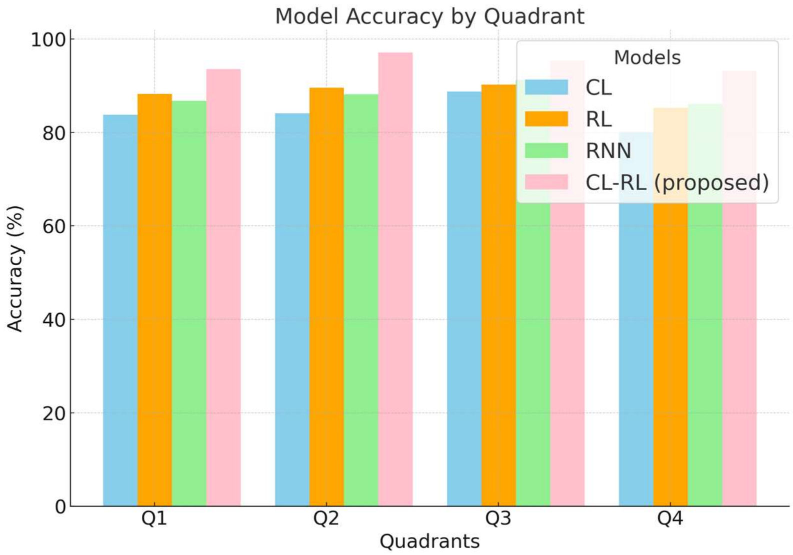

Figure 5. Accuracy of EMOPIA dataset.

6. Conclusion and future work

This work focuses on and contributes to the discipline of Music Emotion Recognition (MER), particularly when evaluating piano performances. This work developed a model that efficiently detects emotional states in piano music by merging Curriculum Learning (CL) with Reinforcement Learning (RL). By leveraging the CL model's capability into RL, this model displays a prominent level of comprehension that reflects how humans think. The recommended CL-RL model addresses the main challenges of MER in piano music: the difficulty and complexity of understanding emotional feelings. Based on comprehensive analysis using the EMOPIA dataset, it is evident that the proposed model outperforms conventional models. The experiment analysis shows that compared to other CL, RL, and RNN models, the CL-RL model significantly improves upon them regarding accuracy, MAE, RMSE, and R^2. The outcome proves that this research work is all-encompassing, especially when conveying the complex feelings of piano performances.

In the future, this work hopes to expand and refine this work for use in many different areas, including Artificial Intelligence (AI), music interpretation, and emotional analysis.

Author contributions

Conceptualization, DZ; methodology, DZ; software, YZ; validation, YZ; formal analysis, DZ; investigation, DC; resources, YZ; writing—original draft preparation, DZ; writing—review and editing, DC; visualization, DC; supervision, DC; project administration, YZ; funding acquisition, DZ. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Cui Y. Vocal music performance evaluation system based on neural network and its application in piano teaching. Revista Ibérica de Sistemas e Tecnologias de Informação. 2023, E55: 451-464.
2. Chang X, Peng L. Evaluation Strategy of the Piano Performance by the Deep Learning Long Short-Term Memory Network. Wireless Communications and Mobile Computing. 2022, 2022: 1-10. doi: 10.1155/2022/6727429
3. Wang W, Pan J, Yi H, et al. Audio-Based Piano Performance Evaluation for Beginners With Convolutional Neural Network and Attention Mechanism. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021, 29: 1119-1133. doi: 10.1109/taslp.2021.3061267
4. Rajesh S, Nalini NJ. Musical instrument emotion recognition using deep recurrent neural network. Procedia Computer Science. 2020, 167: 16-25. doi: 10.1016/j.procs.2020.03.178
5. Chen Q. Intelligent system of piano performance evaluation framework based on multi-dimensional audio recognition algorithm. In: Proceedings of the 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI); 28–30 April 2022; Tirunelveli, India. pp. 82–85. doi: 10.1109/icoei53556.2022.9777178
6. Yu-Chun M, Koong LHC. A study of the affective tutoring system for music appreciation curriculum at the junior high school level. In: Proceedings of the 2016 International Conference on Educational Innovation through Technology (EITT); 22–24 September 2016; Tainan, Taiwan. pp. 204–207. doi: 10.1109/eitt.2016.47
7. Zhang Z, Han J, Coutinho E, et al. Dynamic Difficulty Awareness Training for Continuous Emotion Prediction. IEEE Transactions on Multimedia. 2019, 21(5): 1289-1301. doi: 10.1109/tmm.2018.2871949
8. Wang Y, Sun S. Emotion recognition for internet music by multiple classifiers. In: Proceedings of the 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS); 17–19 June 2019; Beijing, China. pp. 262–265. doi: 10.1109/icis46139.2019.8940288
9. Yang S, Reed CN, Chew E, et al. Examining Emotion Perception Agreement in Live Music Performance. IEEE Transactions on Affective Computing. 2023, 14(2): 1442-1460. doi: 10.1109/taffc.2021.3093787
10. Gao Z, Qiu L, Qi P, Sun Y. A novel music emotion recognition model for scratch-generated music. In: Proceedings of the 2020 International Wireless Communications and Mobile Computing (IWCMC); 15–19 June 2020; Limassol, Cyprus. pp. 1794–1799. doi: 10.1109/iwcmc48107.2020.9148471

11. Zhang K, Wu X, Tang R, et al. The JinYue database for huqin music emotion, scene and imagery recognition. In: Proceedings of the 2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR); 8–10 September 2021; Tokyo, Japan. pp. 314–319. doi: 10.1109/mipr51284.2021.00059
12. Du P, Li X, Gao Y. Dynamic music emotion recognition based on CNN-BiLSTM. In: Proceedings of the 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC); 12–14 June 2020; Chongqing, China. pp. 1372–1376. doi: 10.1109/itoec49072.2020.9141729
13. Wang H, Zhong W, Ma L, et al. Emotional quality evaluation for generated music based on emotion recognition model. In: Proceedings of the 2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW); 18–22 July; Taipei City, Taiwan. pp. 1–6. doi: 10.1109/icmew56448.2022.9859459
14. Kumar S, Rani S, Jain A, et al. Face Spoofing, Age, Gender and Facial Expression Recognition Using Advance Neural Network Architecture-Based Biometric System. Sensors. 2022, 22(14): 5160. doi: 10.3390/s22145160
15. Alnuaim AA, Zakariah M, Alhadlaq A, et al. Human-Computer Interaction with Detection of Speaker Emotions Using Convolution Neural Networks. Computational Intelligence and Neuroscience. 2022, 2022: 1-16. doi: 10.1155/2022/7463091
16. Kimmatkar NV, Babu BV. Novel Approach for Emotion Detection and Stabilizing Mental State by Using Machine Learning Techniques. Computers. 2021, 10(3): 37. doi: 10.3390/computers10030037
17. Balajee RM, Mohapatra H, Deepak V, et al. Requirements identification on automated medical care with appropriate machine learning techniques. In: Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT); 20–22 January 2021; Coimbatore, India. pp. 836–840. doi: 10.1109/icict50816.2021.9358683
18. Mannepalli K, Sastry PN, Suman M. Emotion recognition in speech signals using optimization based multi-SVNN classifier. Journal of King Saud University - Computer and Information Sciences. 2022, 34(2): 384-397. doi: 10.1016/j.jksuci.2018.11.012
19. Balajee R. M., Mohapatra H., Deepak V., Babu D. V., Requirements identification on automated medical care with appropriate machine learning techniques. In: Proceedings of the 2021 6th International Conference on Inventive Computation Technologies (ICICT); 20–22 January 2021; Coimbatore, India. pp. 836–840. doi: 10.1109/ICICT50816.2021.9358683
20. Sekhar Ch, Rao MS, Nayani ASK, Bhattacharyya D. Emotion recognition through human conversation using machine learning techniques. In: Bhattacharyya D, Thirupathi Rao N. (editors). Machine Intelligence and Soft Computing: Proceedings of ICMISC 2020. Volume 1280. pp. 113–122. doi: 10.1007/978-981-15-9516-5_10
21. Durga BK, Rajesh V. A ResNet deep learning based facial recognition design for future multimedia applications. Computers and Electrical Engineering. 2022, 104: 108384. doi: 10.1016/j.compeleceng.2022.108384
22. Ashok Kumar PM, Maddala JB, Martin Sagayam K. Enhanced Facial Emotion Recognition by Optimal Descriptor Selection with Neural Network. IETE Journal of Research. 2021, 69(5): 2595-2614. doi: 10.1080/03772063.2021.1902868
23. Bharti SK, Varadhaganapathy S, Gupta RK, et al. Text-Based Emotion Recognition Using Deep Learning Approach. Computational Intelligence and Neuroscience. 2022, 2022: 1-8. doi: 10.1155/2022/2645381
24. Thirumuru R, Gurugubelli K, Vuppala AK. Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition. Digital Signal Processing. 2022, 120: 103293. doi: 10.1016/j.dsp.2021.103293
25. Kumar S, Haq M, Jain A, et al. Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance. Computers, Materials & Continua. 2023, 74(1): 1523-1540. doi: 10.32604/cmc.2023.028631
26. Srinivas PVVS, Mishra P. A novel framework for facial emotion recognition with noisy and de noisy techniques applied in data pre-processing. International Journal of System Assurance Engineering and Management. 2022. doi: 10.1007/s13198-022-01737-8
27. Mishra P, Srinivas PVVS. Facial emotion recognition using deep convolutional neural network and smoothing, mixture filters applied during preprocessing stage. IAES International Journal of Artificial Intelligence (IJ-AI). 2021, 10(4): 889. doi: 10.11591/ijai.v10.i4.pp889-900
28. Setiawan R, Devadass MMV, Rajan R, et al. IoT Based Virtual E-Learning System for Sustainable Development of Smart Cities. Journal of Grid Computing. 2022, 20(3). doi: 10.1007/s10723-022-09616-z
29. Irrinki MK. Learning Through ICT Role of Indian Higher Education Platforms During Pandemic. Library Philosophy and Practice. 2021.
30. Lakshmi AJ, Kumar A, Kumar MS, et al. Artificial intelligence in steering the digital transformation of collaborative technical education. The Journal of High Technology Management Research. 2023, 34(2): 100467. doi: 10.1016/j.hitech.2023.100467
31. Madhavi E, Lavanya Sivapurapu, Vijayakumar Koppula, et al. B. Esther Rani. Developing Learners' English-Speaking Skills using ICT and AI Tools. Journal of Advanced Research in Applied Sciences and Engineering Technology. 2023, 32(2): 142-153. doi: 10.37934/araset.32.2.142153

32. Nanduri VNPSS, Sagiri C, Manasa SSS, et al. A Review of multi-modal speech emotion recognition and various techniques used to solve emotion recognition on speech data. In: Proceedings of the 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA). 3–5 August 2023; Coimbatore, India. pp. 577–582. doi: 10.1109/icirca57980.2023.10220691
33. Srinivas P, Khamar SN, Borusu N, et al. Identification of Facial Emotions in Hitech Modern Era. In: Proceedings of the 2023 2nd International Conference on Edge Computing and Applications (ICECAA); 19–21 July 2023; Namakkal, India. pp. 1202–1208. doi: 10.1109/icecaa58104.2023.10212285
34. Sugumar R, Sharma S, Kiran PBN, et al. Novel method for detection of stress in employees using hybrid deep learning models. In: Proceedings of the 2023 8th International Conference on Communication and Electronics Systems (ICCES); 1–3 June 2023; Coimbatore, India. pp. 984–989. doi: 10.1109/ICCES57224.2023.10192609
35. Kuchibhotla S, Dogga SS, Vinay Thota NVSLG, et al. Depression detection from speech emotions using MFCC based recurrent neural network. In: Proceedings of the 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN). 5–6 May 2023; Vellore, India. pp. 1–5. doi: 10.1109/vitecon58111.2023.10157779
36. Chintalapudi KS, Patan IAK, Sontineni HV, et al. Speech emotion recognition using deep learning. In: Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI); 23–25 January 2023; Coimbatore, India. pp. 1–5. doi: 10.1109/iccci56745.2023.10128612
37. Fernandes Dimlo UM, Bhanarkar P, Jayalakshmi V, et al. Innovative method for face emotion recognition using hybrid deep neural networks. In: Proceedings of the 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI); 11–13 April 2023; Tirunelveli, India. pp. 876–881. doi: 10.1109/icoei56765.2023.10126007
38. Nallapu SK, Boddukuri VB, Ganesh DVVALS, et al. Intelligent video analytics & facial emotion recognition using artificial intelligence. In: Proceedings of the 2023 Second International Conference on Electronics and Renewable Systems (ICEARS); 2–4 March 2023; Tuticorin, India. pp. 896–900. doi: 10.1109/icears56392.2023.10084928