Research on Music Emotion Analysis and Classification Algorithm Based on Artificial Intelligence Natural Language Processing

Ping Ran
College of Humanities and Arts, Xi'an International University, China
Ping_ran176@outlook.com

Abstract: *Emotion detection in music is one of the important features that enhances user experience and interaction in digital media. Experiments were conducted using the Music Information Retrieval Evaluation eXchange (MIREX) multimodal emotion dataset, consisting of 903 tracks, with a balanced class distribution of four emotions: happy, sad, anger, and fear. Traditional methods often cannot represent the emotional characteristics of music properly. It makes the categorization of emotions improper and inaccurate due to its limitations. These challenges have been addressed in the current research by proposing a hybrid model with a new approach towards music emotion analysis through Generalized Autoregressive Pretraining for Language Understanding (XLNet) combined with a redefined Crayfish Optimization (CO) algorithm. The proposed model can detect and classify the emotional information in music effectively, whereby the integration of modern Natural Language Processing (NLP) and advanced optimization methods proves to be effective. The data was split into 70% training, 15% validation, and 15% testing, with 5-fold cross-validation to ensure robust evaluation. The proposed technique has obtained an accuracy of 92.3% ± 0.5%, precision of 93.8%, recall of 91.6%, and F1-score of 94.2%. These results are reported as mean ± Standard Deviation (SD) over 5 cross-validation folds, showing robust statistical confidence. These scores show how the model detected complex emotional variations in music and delivered a great result compared to other traditional approaches. Hence, it would be very useful in applications such as emotion classification, personalized content circulation, and recommendation systems for music.*

Keywords: *Music emotion analysis, redefined crayfish optimization, XLNet, NLP, emotion classification.*

Received July 04, 2025; accepted February 09, 2026
https://doi.org/10.34028/iajit/23/5/8

1. Introduction

The music is an international language wherein human feelings and experiences are involved. The inability to precisely identify and categorize the expressed emotions in music has led to the emergence of a high rise in the usage of the latest technologies towards better comprehension and applicability of its content [14]. Recent advances within the area of Natural Language Processing (NLP) and artificial intelligence opened many fields and gave solid tools for deciphering such complex data [36]. The development transcended conventional techniques by offering new possibilities of analysis of musical emotional content [8]. Bidirectional encoder representations from transformers-like models perform outstandingly on most tasks related to the interpretation of natural language [26]. It comprises pre-processing of lyrics, audio features, and metadata, to which Artificial Intelligence (AI) has to be integrated with music emotion analysis [35]. The ultimate vision lies in coming up with algorithms that would efficiently decode and classify the emotional content of a musical piece [31]. Meta-learning, most commonly known as "learning to learn," plays an important role in optimizing these models [15].

It helps to develop the processes of learning on diverse datasets and improves general accuracy and adaptability of AI algorithms [6]. Recent research tries to identify patterns and associations of music and emotions, thereby enhancing applications in areas like music recommendation and entertainment [12]. The use of models like Bidirectional Encoder Representations from Transformers (BERT) and adapting those into meta-learning methods will enhance their power in classifying emotions. It has to incorporate complex models with sophisticated optimization approaches, which are able to overcome many drawbacks of traditional approaches [34]. Applications enabled by these systems will be more functional and accurate as emotion detection and classification in music is bound to rise [19]. Enhanced user experience, better support for good mental health, and creative interactive applications are some obvious benefits of enhanced emotion recognition [17]. These studies would therefore have a great impact on domains like emotion recognition and those concerned with music information retrieval. It aims to set new benchmarks in the interpretation and utilization of the emotional content of music, incorporating AI and NLP techniques [32].

Based on modern AI and NLP techniques, the present work proposes a hybrid model called Redefined Crayfish Optimization-driven XLNet (RCO-XLNet) for effective

music emotion analysis and classification.

2. Related Works

Rajesh and Nalini [25], by deep learning methods, developed a deep recurrent neural network that identified the emotion by the classification of musical devices. The results showed that in the case of musical emotion detection from musical clips, it was much better when deep learning-based Recurrent Neural Networks (RNNs) are employed.

Sarkar *et al*. [28] proposed a deep Convolutional Neural Network (CNN) for recognizing musical emotions. Their proposed CNN was based on the improved version of Visual Geometry Group Network (VGGNet) but with considerably fewer numbers of layers. Experimental results showed that the suggested network remarkably increased the efficiency of the recognition.

Gudivaka [10] explored the role of AI and Big Data in music education for personalized and interactive learning experiences. A similar approach is used in the present work, making use of AI algorithms and NLP techniques to analyze musical features and lyrics in order to classify emotions. These advantages consist of better emotional comprehension, feedback on the go, and customization which all result in having more engaged learners and learners with better learning outcomes.

Sheykhivand *et al*. [29] developed a new approach that was based on the identification of musical emotions with the integration between Long Short-Term Memory (LSTM) and CNN. When both networks were integrated together, then the accuracy and stability for the performance of the system were enhanced. The suggested approach was compared with some similar procedures, where excellent outcomes appeared.

Chen and Li [2], in the area of music emotion recognition, proposed a hybrid network including CNN-LSTM. They effectively reflected the emotions on variable models by using the stacking-based multimodal collaborative learning approach. Its effect of classification was outstanding when compared to single-modal classification.

The work of Pandeya *et al*. [24] proposed a helpful affective computing system to carry out the analysis of the emotional state of a subject. Many unimodal and multimodal architectures were developed for evaluating from scratch video, audio, and facial emotions. Finally, boosting approaches together with the exchange of information from audio to video allowed them to lower the computing expenses.

Thus, Pandeya and Lee [23] utilized the CNN-based multimodal architecture for classifying a wide range of human emotions depicted in music clips. They generated a rich set of music clip emotions based on several factors, such as languages, cultures, and musical styles. The results proved that the CNN-based multimodal network proposed could classify high-level human emotions correctly.

Deng and Lin [5] suggested an effective algorithm for music sentiment analysis called Deep Reinforcement Learning (DRL). Indeed, it had an excellent performance in the classification tasks of emotional responses. In their experimental results, the proposed technique outperformed conventional methods significantly.

Deng *et al*. [4] suggested a framework that uses Contrastive Language-Image Pretraining (CLIP) to classify Image Emotion Classification (IEC), and proposed a prompt tuning method, which approximates the pretraining goal of CLIP and automatically generates instance-specific prompts by conditioning them on the categories and image contents of instances, and diversifies the prompts to prevent suboptimal problems. The given strategy is followed in the proposed work, where the musical characteristics and lyrics are analyzed with the help of AI algorithms and NLP to classify emotions, which improves the model in recognizing emotional content in music. Its advantages are that it is more accurate in the classification of emotions, generalizes better across genres of music, and more refined in its perception of emotional expression in music.

By contrast, Fong *et al*. [7] proposed a theory-driven explainable deep learning CNN classification system to predict the musically time-varying emotional response. Furthermore, the authors introduced a new filter for CNN using the structure of frequency harmonics. The proposed model outperformed other models and turned out to be cost-effective.

Based on the detection of features, moods, and feelings of a specific user, Moscato *et al*. [20] proposed an innovative approach to music recommendation. Content-based filtering models were developed to be more accurate and effective by including the personality and mood of the user.

Hizlisoy *et al*. [13] proposed a technique for musical emotion classification by using a convolutional Long Short-Term Memory (LSTM) deep neural network system. The results showed that the identification performance would get better if new features were added to the baseline audio features. This technique produces more accurate results.

Shi *et al*. [30] presented a single generative model of image emotion classification which combines various emotion models and represents intricate semantic associations among emotion labels. The given method is followed in the proposed work, which uses a flexible natural language template and incorporates multimodal pre-trained models to improve emotion classification in music. The advantages are better accuracy, flexibility in a variety of musical styles, and a more sophisticated perception of emotional manifestations in music.

Grekow [9] used RNN to test the ability in detecting emotions from musical segments. The emotion detection was done as continuous values using trained regression algorithms. Several experiments were conducted on data

with variable feature sets and segmentations, while the models were developed using the WekaDeeplearning4j package. Experimental results of the research could help in the building of automated systems for recognizing music emotions.

Naser and Saha [21] proposed a data processing-based system on pattern identification for identifying the emotional impact of musical preference. The brain pattern due to arousal and dominance was more discriminative, comparing low-rated music videos with high-rated ones. The technique worked fantastically in categorizing emotions.

Wang [33] presented the analysis of the neural network algorithm and musical gadgets with a view to enabling emotion recognition and composition. A new architecture for a music composition neural network has been developed to extend the modification of the LSTM generation network. That was pretty good, and it might lead to a leap in improving the model's accuracy in music composition. Table 1 compares various approaches to music emotion analysis based on supervision, dataset, model family, modality, and reported performance.

Table 1. Comparison of approaches for music emotion analysis.

Criteria	Supervision	Dataset(s)	Model Family	Modality	Accuracy (%)
Rajesh and Nalini [25]	Supervised	Audio clips	RNN	Audio	85
Sarkar *et al*. [28]	Supervised	Audio clips	CNN	Audio	86.5
Sheykhivand *et al*. [29]	Supervised	EEG signals & Audio clips	LSTM + CNN	EEG + Audio	87
Chen and Li [2]	Supervised	Multimodal (audio + lyrics)	CNN + LSTM	Audio + Lyrics	88
Pandeya and Lee [23]	Supervised	Music clips	CNN-based multimodal network	Video, Audio, and Facial	89
Proposed RCO-XLNet	Supervised	Multimodal (MIREX)	XLNet + RCO	Audio + Lyrics	92.3

2.1. Problem Statement

Although Multimodal Music Emotion Recognition (MMER) has been developed considerably, there are still difficulties associated with the use of audio and text modalities and their efficient integration. Most of the models use the methods of static fusion, which could fail to reflect the dynamism of modalities interaction [18]. Moreover, hyperparameter optimization on these modalities has not been properly studied, and thus, this factor may restrict the output of the MMER systems [27].

In order to solve these problems, this paper suggests Redefined Crayfish Optimization-driven XLNet (RCO-XLNet) model that uses XLNet autoregressive pretrain characteristics to analyze text and implements the optimization methods that improve multimodal fusion. Redefined Crayfish Optimization-driven XLNet (RCO-XLNet) should enhance the accuracy and generalizability of MMER systems by concentrating on dynamic fusion and robust optimization.

3. Methodology

First, the data is collected from the multi-modal MIREX emotion dataset and then it is preprocessed using methods like tokenization, lowercase conversion and stemming. After preprocessing, methods like Mel-Frequency Cepstral Coefficient (MFCC) and Term Frequency Inverse Document Frequency (TF-IDF) are used for feature extraction. After extraction, Lyric TF-IDF features and audio MFCC features are concatenated into one feature vector. This basic concatenation method allows the model to handle both features at the same time. The merged characteristics are transmitted via a common trunk network, in which the textual and audio data are handled simultaneously and then classified. The training is controlled using categorical cross-entropy loss, which is suitable in multi-class emotion classification problems. The overall flow of the system is depicted in Figure 1. Figure 2 displays the architecture for a multi-modal music emotion analysis pipeline.

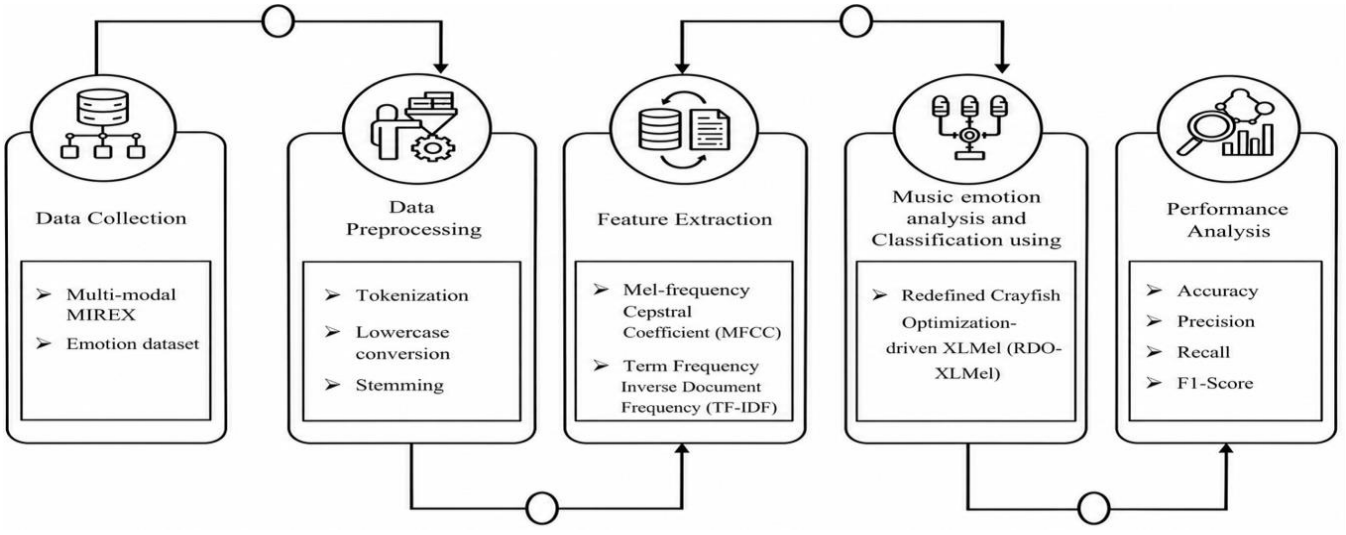


Figure 1. Basic structure of the system.

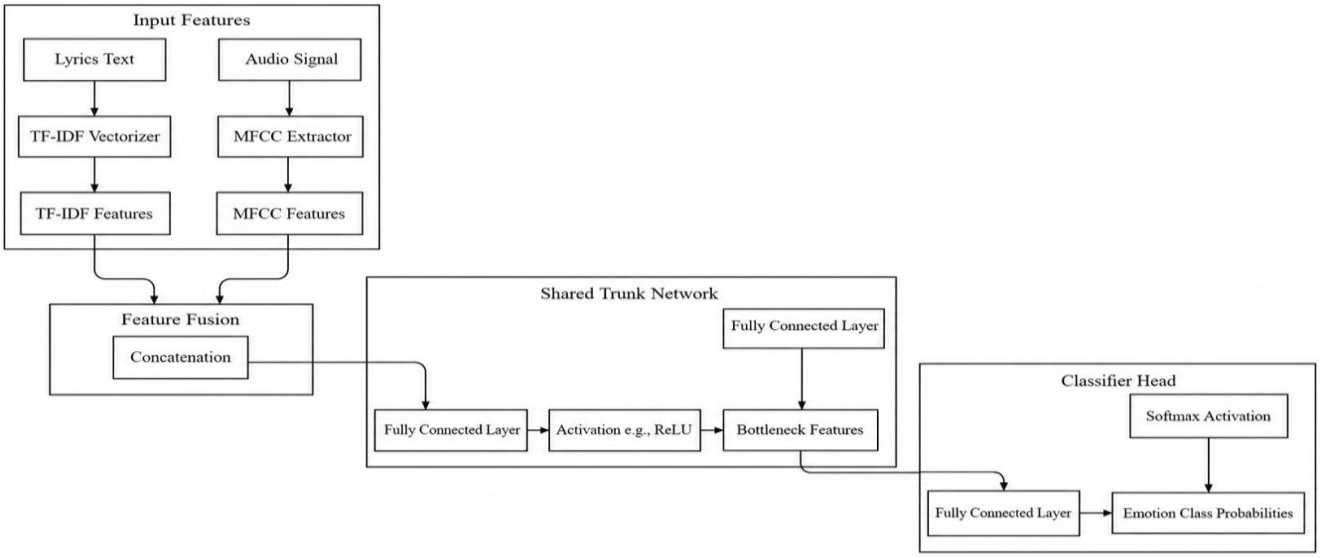


Figure 2. Architecture for a multi-modal music emotion analysis pipeline.

3.1. Dataset

The multimodal MIREX emotion dataset encompasses a wide variety of parameters in music, lyrics, and metadata for the task of musical emotion identification. This dataset would be ideal to train models using modern AI and NLP techniques that categorize and predict emotional responses in music, considering that it contains pre-extracted features, including MFCCs and chroma, along with the emotional labels of each track. The dataset includes approximately 903 songs which are categorized into four categories of emotions, happy, sad, anger and fear with a duration of about 2.5 to 6.5 minutes. Stratified sampling is used to keep the balance of classes by dividing it into 70% training, 15% validation and 15% testing sets. Each track contains lyrics; where the lyrics were not available, zero-padding was used to maintain the text input format. The data is associated with the official MIREX 2013 Audio and Lyrics-based Music Emotion Recognition Task and is available on Kaggle under Creative Commons Attribution-NonCommercial 4.0 License, which permits its use in non-commercial research with proper citation [22]. Source: Kaggle (https://www.kaggle.com/datasets/imsparsh/multimodal-mirex-emotion-dataset).

3.2. Preprocessing

In text mining and Natural Language Processing (NLP), preprocessing is a crucial stage and most important activity. For text documents, preprocessing methods including tokenization, lowercase conversion, and stemming are performed.

- **Tokenization:** It is employed to find the important keywords. Tokenization difficulties vary depending on the language. Since most words in languages like French and English are separated from one another by white spaces, these languages are known as space delimited languages. The writing system and word typography have an impact on tokenization.
- **Lowercase conversion:** It is applied to words that have an identical appearance each time. Since all of the text in this method is written in lowercase, content analysis is quite easy.
- **Stemming:** It is applied to determine a word's root or stem. This method aims to save time and memory by removing different suffixes, and creating perfectly matched stems. To minimize the size of the text document, two different kinds of stemming algorithms are employed. They are affix removal stemmers and successor varieties.
- **Affix Removal Stemmers:** Affix removal stemmers eliminate terms' prefixes or suffixes before they leave the stem. The following rule will apply to an affix removal stemmer:

If a word ends with "s" and not "us" or "ss", Then "s" -> "null", Example: Books - Book.

- **Successor Variety:** The number of letters that come after a string in a word within a body of text is called as successor variety. For example, the successor variety of the words stemming and stemmer is identified by counting the number of different characters that follow each substring inside the terms. In stemming, the characters "m," "i," "n," and "g" come after the substring "Stem," whereas in "Stemmer," "m," "e," and "r" come after the substring. Users can recognize morpheme boundaries by contrasting these successor variations, which aids in separating the base word "stem" from its suffixes.

By using these effective preprocessing approaches, noise is removed from text input and actual words may be identified by their root word.

Whereas XLNet is able to capture the context of the lyrics, TF-IDF is able to identify the important terms that help in classifying emotions. This two-fold strategy will make sure that not only significant individual words but also the context on the whole are considered in the

model.

3.3. Feature Extraction

Mel-Frequency Cepstral Coefficient (MFCC) is employed for feature extraction in audio data to capture spectral properties, while Term Frequency Inverse Document Frequency (TF-IDF) is applied to text data for accurately analyzing the emotions. XLNet representations give rich contextual embeddings of the lyrics, which reflect the rich relationships between words, whereas TF-IDF features emphasize the significance of particular words in the song. The combination of the two gives the model the advantages of the semantic knowledge of XLNet and the term significance of TF-IDF.

3.3.1. Mel-Frequency Cepstral Coefficient (MFCC)

The Mel-Frequency Cepstral Coefficient (MFCC) computation is performed step-by-step as follows:

- **Pre-Emphasis:** As indicated by Equation (1), the voice signal $w(m)$ must be sent over a high-pass filter.

$$z(m) = w(m) - b * w(m-1) \quad (1)$$

When the output wave is denoted by $z(m)$, and the value of b typically ranges from 0.9 to 1.0. The signal is processed by running it through a filter that highlights its higher frequency. At higher frequencies, the signal's energy goes up.

- **Frame Blocking:** The audio signal is split up into frames, which are 20-30 *ms* long segments. Every frame that is next to another frame may be separated by $N(N<M)$ for up to M frames. In general, N and M are set at 100 and 256, respectively.
- **Windowing:** This method is employed for segmenting an audio stream into temporal sections in signal processing applications. The signal's repeating moves have nothing to do with the signal that exists in reality. The smooth function known as the windowing function has extreme values of zero. The discontinuity at the boundaries is supposed to be hidden. *A* in the signal occurs following the window function's application, although the impact on signal statistics is not so severe.

MFCC typically uses the Hamming Window function as its window function. The frame's initial and final points are kept continuous. Equation (2) defines the hamming window function; the window function is represented by $X(m)$.

$$Z(m) = W(m) * X(m) \quad (2)$$

- **Fast Fourier Transform:** It is a signal analysis approach that makes audio processing easier by extracting and compressing some aspects of the audio signal without missing any important information. It provides a frequency domain representation of the given signal. All necessary information from the original signal remains unchanged, only when the signal's representation is altered.
- **Discrete Cosine Transform (DCT):** Equation (3) provides the formula for the discrete cosine transform.

$$D(m) = \sum_{k=0}^{K-1} E_k \cos\left(m \cdot (l-0.5) \cdot \frac{\pi}{40}\right) \quad (3)$$

Where m=0,1….. *to* M and M is the number of triangular band pass filters. Mel-scale coefficients are produced by performing the DCT on the band pass filter output.

3.3.2. Term Frequency Inverse Document Frequency (TF-IDF)

After presenting the mathematical framework, its behavior is discussed with respect to each variable; later, the structure and implementation of TF-IDF will be delivered for a set of texts.

- **Mathematical Framework:** The inverse distribution of a word across the entire document sample is compared to the relative frequency of a particular word in a document to determine how TF-IDF operates. It makes clear sense that this computation establishes a word's significance within a specific document. The general process for implementing TF-IDF is as follows, with a few small variations. Equation (4) shows how to calculate x_c given a word x, a document collection C, and a single document $c \in C$.

$$x_c = e_{x,c} * \log\left(\frac{|C|}{e_{x,C}}\right) \quad (4)$$

Where $|C|$ is the corpus size, $e_{x,c}$ is the total number of cases in which$x$ occurs in c, and $e_{x,C}$ is the total number of cases in which x occurs in C. Various scenarios can arise for every word in this case, based on the values of $e_{x,c}$, $|C|$, and $e_{x,C}$.

Assume that the length of the corpus is about the same as the frequency of x over C, or $|D| \sim fw$, D. If x_c is smaller than $e_{x,c}$ but still positive, then $J < \log\left(\frac{|C|}{e_{x,C}}\right) < d$ for some small constant d. This suggests that while x is comparatively prevalent across the corpus, it maintains some significance throughout C. If $e_{x,c}$ is large and $e_{x,C}$ is small, then x_c will also be large since $\log\left(\frac{|C|}{e_{x,C}}\right)$ will be fairly large.

- **Encoding TF-IDF:** The basic nature of the TF-IDF code makes it simple. Given a query r consisting of a collection of words x_j, the $x_{j,c}$ will be computed for each document $c \in C$ that contains x_j. The most basic way to accomplish this is to go through the collection of documents and maintain a continuous total of $e_{x,c}$ and $e_{x,C}$. After that, $x_{j,c}$ is simple to estimate.

Equation (5) states that the documents are returned in a decreasing sequence. This is how TF-IDF is traditionally implemented.

$$\sum_j x_{j,c} \quad (5)$$

3.4. Redefined Crayfish Optimization-driven XLNet (RCO-XLNet)

The new hybrid model Redefined Crayfish Optimization-driven XLNet (RCO-XLNet) improves music emotion analysis and classification by incorporating an improved variant of the Crayfish Optimization (CO) algorithm into the XLNet language model. CO algorithm, inspired by the behavior of crayfish, is an exceptional tool to explore and utilize sophisticated optimization settings; hence, it is perfect for optimizing the hyperparameters of XLNet. The XLNet model, with a permutation-based generation and context awareness from both directions, has an excellent basis for musical emotion understanding and categorization from text and audio. XLNet is not a frozen feature extractor but is specifically trained on lyrics to extract the emotional context of textual data. During training, the pre-trained representations in the model are modified to maximize music emotion classification performance. This range of multidimensional parameters of XLNet is dealt successfully by RCO in an attempt to improve the model's performance over standard crayfish techniques.

This hybrid method makes use of the flexibility that RCO provides in enhancing the detection and classification capability of XLNet on emotional content of music, eventually leading to an accurate and effective emotion identification system. Combining both methods will give the advantages of optimization and NLP together to produce the best outcomes in music emotion analysis. Once the XLNet and TF-IDF features have been extracted, they are concatenated and sent through a fully connected classifier head, which makes the prediction of the emotional class of the music. The classifier head is a series of dense layers with a softmax activation to provide the likelihood of emotions. The algorithm for the proposed method is shown in Algorithm 1.

Algorithm 1. *Pseudocode for RCO-XLNet: Crayfish Optimization-driven XLNet for Music Emotion Recognition*

```
Step 1: Initialize Dataset
data = load_data("path_to_data")
# Step 2: Preprocess Data
# Text Data: Tokenization, Lowercasing, Stemming, TF-IDF
# Audio Data: Extract MFCC Features
text_data, audio_data = data_preprocessing.preprocess(data)
text_features = feature_extraction.tfidf(text_data)
audio_features = feature_extraction.mfcc(audio_data)
# Step 3: Initialize XLNet Model
model = xlnet.initialize_model()
# Step 4: Define the Optimization Process with Crayfish Optimization (RCO)
def redefined_crayfish_optimization(model, features, parameters):
    # Initialize Crayfish Optimization
    crayfish_opt = crayfish_optimization.initialize(parameters)

    # Iterating until convergence
    while not convergence_criteria_met():
        # Evaluate performance of the model
        performance = model.evaluate(features)

        # Update positions using Crayfish Optimization
        crayfish_opt.update_positions(performance)

        # Update model parameters based on the best solution found
        model.update_parameters(crayfish_opt.best_solution)

    return model

# Step 5: Optimize Model with RCO
optimized_model = redefined_crayfish_optimization(model, text_features + audio_features, optimization_parameters)
# Step 6: Train the Model with Optimized Hyperparameters
optimized_model.train(text_features, audio_features)
# Step 7: Evaluate the Model Performance
performance = optimized_model.evaluate(text_features, audio_features)
print("Model Performance: ," performance)
```

In Figure 3, the flowchart of the RCO-XLNet algorithm is shown, giving an overview of the different steps from data loading and preprocessing to model training and evaluation.

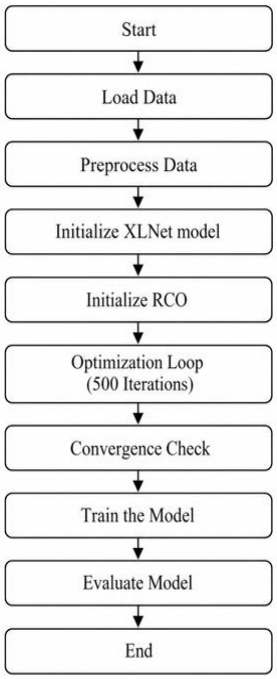


Fig. 3. RCO-XLNet Algorithm Flowchart.

3.4.1. XLNet

XLNet is an effective language model that focuses on permutation-based training, capturing bidirectional information for better interpretation of texts. It uses rich contextual data and very strong correlations in textual data for efficient comprehension and contextualization of the lyrics, metadata, and audio aspects of music emotion analysis. Thus, it provides an accurate result in the classification of emotions. Let the input series with length S be denoted by [1, 2, 3,…, S]. The permutation language model has the mathematical formulation, as shown in Equation (6).

$$\max_{\theta} F_{t-T_S}\left[\sum_{s=0}^{S} \log o_\theta(y_{t_s}|Y_{t<s}\right] \quad (6)$$

In general terms, XLNet employs two-stream attention mechanisms (content stream and query stream) to accomplish permutation modeling in the transformer framework. Equation (7) illustrates how conventional hidden conditions are represented by the content stream attention. In Equation (8), the position signal is represented by the query stream attention.

$$g_{t_s}^{(b)} \leftarrow Attn_{layer}\left(R = g_{t_s}^{(b-1)}, LU = g_{t\leq s}^{(b-1)}; \theta\right) \quad (7)$$

$$h_{t_s}^{(b)} \leftarrow Attn_{layer}\left(R = h_{t_s}^{(b-1)}, LU = g_{t<s}^{(b-1)}; \theta\right) \quad (8)$$

Where the number of self-attention levels is indicated by the sign *b*. Query, key, and value are represented by the symbols *R*, *L*, and *U*. Additionally, XLNet slows the self-regressive model's progress by using partial estimation to lessen the optimization difficulty. Equation (9), which is based on a non-target subsequence ($t_{\leq d}$) indicates that the goal of the issue is to maximize the output or target subsequence ($t_{>c}$). For both target and non-target subsequences, the cutting point is indicated by *d* in the specified factorization order (*t*).

$$\max_{\theta} F_{t-T_S}\left[\sum_{s=0}^{S} \log o_\theta(y_{t>d}|T_{t\leq d}\right] = F_{t-T_S}\left[\sum_{s=d+1}^{S} \log o_\theta(y_{t_s}|T_{t<s}\right] \quad (9)$$

Moreover, XLNet integrates two key Transformer-XL techniques: segment recurring algorithm and comparative positional encoding. These methods enhance the input sequence's long-term dependency. For this reason, the XLNet model works well for NLP tasks.

3.4.2. Redefined Crayfish Optimization (RCO)

RCO extends the CO algorithm with more sophisticated exploration and exploitation techniques in high-dimensional spaces. A population size of 50 crayfish agents was utilized in the optimization process, and the process was repeated 500 times. The training had an early-stopping criterion which was based on validation performance and the process was stopped when no improvement was realized in 10 consecutive generations. To guarantee the reproducibility of results, a random seed of 42 was used. The hyperparameter search space consisted of the following values: learning rate 0.001-0.1, batch size 16, 32 or 64, XLNet layers unfreezing 6 to 12 layers, max sequence length 128 or 256, dropout 0.1-0.5, AdamW optimizer, weight decay 0.01-0.1 and class weights to balance classes (happy, sad, anger and fear).

The slow converging time and easy local area fallout are two of the CO algorithm's drawbacks. As a result, the experiment employs four techniques to increase it and fully optimize the CO algorithm's performance. To make the starting population more uniformly dispersed throughout the search space, the Halton sequence is initially employed to enhance population initialization. Second, a greedy technique is used to pick the next generation of the population, expanding the search place and improving the variety of the crayfish population. This strategy generates the quasi-oppositional solution of the generation. Thirdly, in an effort to increase the foraging stage's pace of optimization, the robust guiding factor is added. The marine predator method's vortex effect is finally included after the foraging stage to improve the algorithm's capacity to break out of the local optimal.

- **Halton Sequence Population Initialization:** This procedure serves as the foundation for algorithm iteration. Crayfish populations are created at random by a typical CO algorithm. The population cannot cover the entire search space with this initialization method, which can lead to an unequal distribution of small and medium-sized crayfish. This can cause the algorithm to converge slowly or even prematurely into a local optimal. Indeed, it is a straightforward and simple method of implementation. Because of its deterministic character, it is not like random sequences that produce different spots every time.

Equation (10) displays the formula for the starting position of every crayfish created by population initialization using the Halton sequence.

$$W_{j,i} = ka_i + (va_i - ka_i).Haltonset, j = 1,2,\ldots.,M; i = 1,2,\ldots.,C \quad (10)$$

Where the Halton sequence is used to determine the value of Haltonset.

- **Quasi Opposition-Based Learning (QOBL):** The idea is to produce the opposing solution to the one that the current population has, compare it to the opposite solution, and choose the superior candidate solution to be the population's future generation. Equation (11) displays the position representation of the opposition solutions that were created.

$$W_j^{OBL} = (ka + va) - W_j \quad (11)$$

Where the problem variable's lower bound (*ka*) and upper bound (*va*) are expressed in the search place. Equation (12) illustrates the value of *N*. It necessitates that the opposing solution remain near the center *N* of the search place rather than the opposite value. Equation (13) displays W_j^{OBL}, the quasi-opposite solution generated by QOBL.

$$N = \frac{ka + va}{2} \quad (12)$$

$$W_j^{QOBL} = \begin{cases} N + \left(W_j^{OBL} - N\right).q, W_j < N \\ W_j^{OBL} + \left(N - W_j^{OBL}\right).q, else \end{cases} \quad (13)$$

Where *q* is anequally distributed random value that falls between 0 and 1.

- **Elite Steering Factor:** Equation (14) illustrates the way it is utilized to control the position update. If the crayfish location updated by Equation (14) is out of the search boundary, the boundary value is taken.

$$W_{j,i}^{s+1} = O.\left(W_{j,i}^{s} - W_{food}\right) + O.q.W_K \quad (14)$$

Where the elite is represented by W_K, the current crayfish population's ideal location.

- **Fish Aggregation Device Effect:** The CO algorithm incorporates the Fish Aggregation Device (FAD) effect into the predation phase. In the event that the crayfish $fitness_j^{s+1}$ following the update of Equation (14) is not as high as the $fitness_j^s$ of the preceding generation, the FAD effect will cause the crayfish places to be affected. Equation (15) provides a mathematical expression for this process.

$$W_j^{s+1} = \begin{cases} W_j^s + CF.\left(ka + q.(va - ka)\right).V, q \le FADs \\ W_j^s + \left(FADs.(1-q).\left(W_{q1}^s - W_{q2}^s\right)\right), q > FADs \end{cases} \quad (15)$$

Where Fish Aggregation Devices (FADs) is equal to 0.2, V is a vector with a value of 0 to 1, q is an equally distributed random value that belongs to [0, 1], W_j^s is the current location of the j crayfish, $W_j^{s+1}$is the location of the next generation, and W_{q1}^s and W_{q2}^s are the locations of two random members of the present generation.

4. Result

The proposed approach is implemented using Python (v 3.12) on Windows 11 OS. The system is driven by an Intel Core i7 processor and features a high-performance IRIS graphics card, which delivers a strong capacity for executing machine learning applications.

Here, the proposed method Redefined Crayfish Optimization-driven XLNet (RCO-XLNet) is compared with existing methods such as Extreme Gradient Boosting (XGB) Classifier, Random Forest (RF), and Support Vector Machine (SVM) Linear (Gujar [11]) using the metrics like precision (%), f1-score (%), accuracy (%) and recall (%). Figure 4 shows the confusion matrix for emotion classification, which includes different types of emotions like happy, sad, anger, and fear. Out of 903 audio files, 843 are correctly detected. This shows the efficiency of the model in recognizing and categorizing musical emotions.

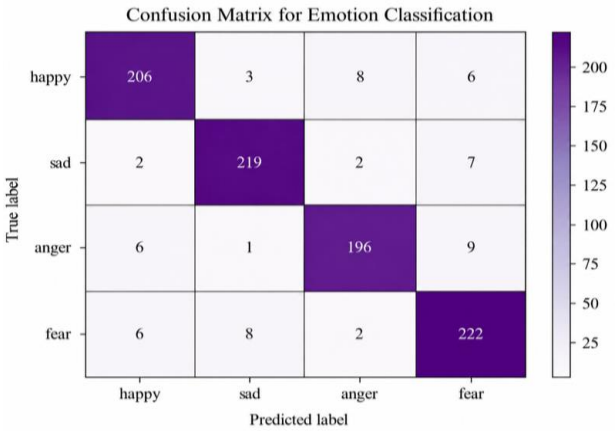


Figure 4. Confusion matrix for emotion classification.

4.1. Accuracy

It gives the ratio of correctly classified cases out of all instances. High accuracy means that the model tends to classify most of the time correctly during music emotion analysis. The output of accuracy is shown in Figure 5 and Table 2. The result demonstrates that our proposed RCO-XLNet (92.3%) technique performs better than the existing methods such as RF (86.19%), XGB Classifier (86.3%), and SVM Linear (85.45%).

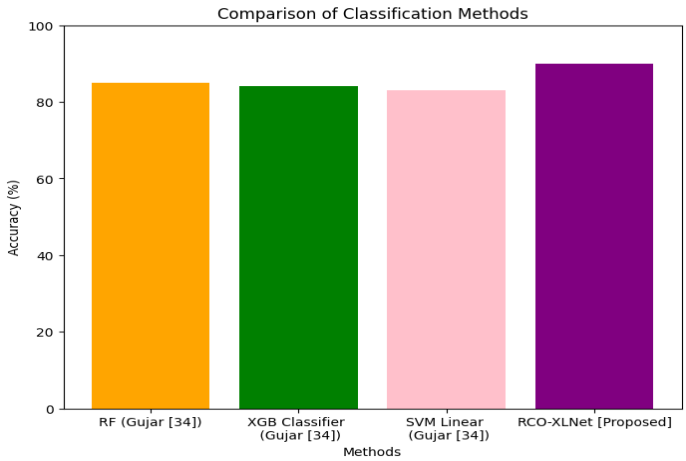


Figure 5. Accuracy.

Table 2. Comparison of Accuracy.

Methods	Accuracy (%)
RF (Gujar [11])	86.19
XGB Classifier (Gujar [11])	86.3
SVM Linear (Gujar [11])	85.45
RCO-XLNet [Proposed]	92.3

4.2. Precision

It is conventionally defined as a ratio of true positives against the sum of false and true positives. A high level of precision in identifying emotions in music indicates that the model can reduce false positives, thereby predicting the emotions correctly. The output of precision is depicted in Figure 6 and Table 3. The result shows that our proposed RCO-XLNet (93.8%) approach outperforms the existing approaches such as RF (86.195%), XGB Classifier (86%), and SVM Linear (86.9%).

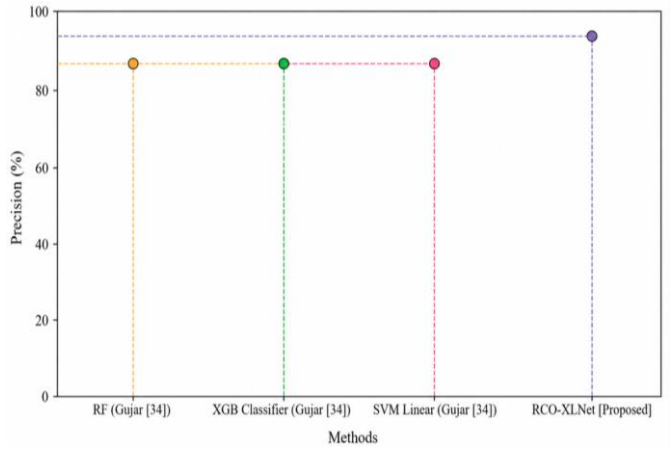


Figure 6. Precision.

Table 3. Comparison of precision.

Methods	Precision (%)
RF (Gujar [11])	86.195
XGB Classifier (Gujar [11])	86
SVM Linear (Gujar [11])	86.9
RCO-XLNet [Proposed]	93.8

4.3. Recall

It is calculated by dividing the number of true positives by the total number of false negatives and true positives. In the field of music emotion detection, high recall means the model found most occurrences of an emotion. The output of recall is shown in Figure 7 and Table 4. The result indicates that our proposed RCO-XLNet (91.6%) technique performs better when compared to existing methods such as RF (86%), XGB Classifier (86.3%), and SVM Linear (46.04%).

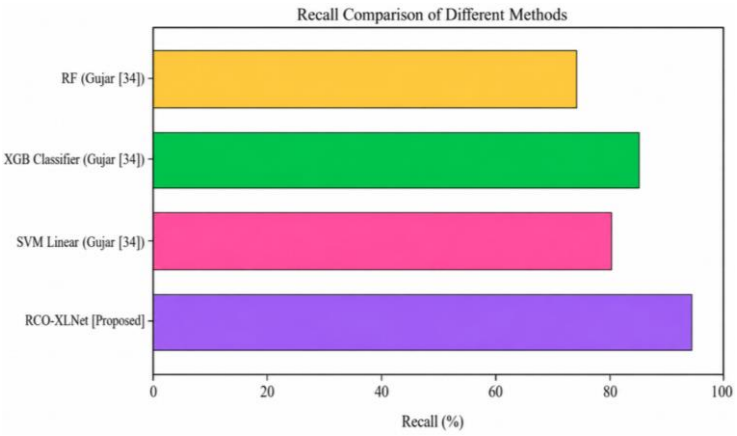


Figure 7. Recall.

Table 4. Comparison of recall.

Methods	Recall (%)
RF (Gujar [11])	86
XGB Classifier (Gujar [11])	86.3
SVM Linear (Gujar [11])	46.04
RCO-XLNet [Proposed]	91.6

4.4. F1-Score

It is a type of balanced measure and the harmonic mean of recall and precision. A high F1-score in the classification of emotions in music will indicate that a model is getting a good balance between recall and precision, hence making enhancements in the overall classification performance. The output of f1-score is depicted in Figure 8 and Table 5. The result demonstrates that our proposed RCO-XLNet (94.2%) technique outperforms the existing methods such as RF

(86.1%), XGB Classifier (86%), and SVM Linear (83.56%).

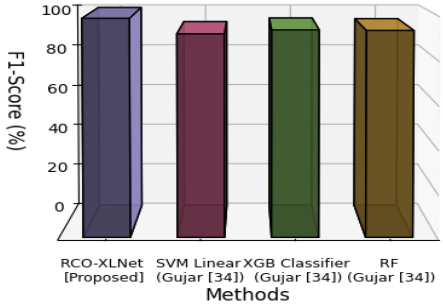


Figure 8. F1-score.

Table 5. Comparison of F1-score.

Methods	F1-Score (%)
RF (Gujar [11])	86.1
XGB Classifier (Gujar [11])	86
SVM Linear (Gujar [11])	83.56
RCO-XLNet [Proposed]	94.2

4.5. Comparative Analysis

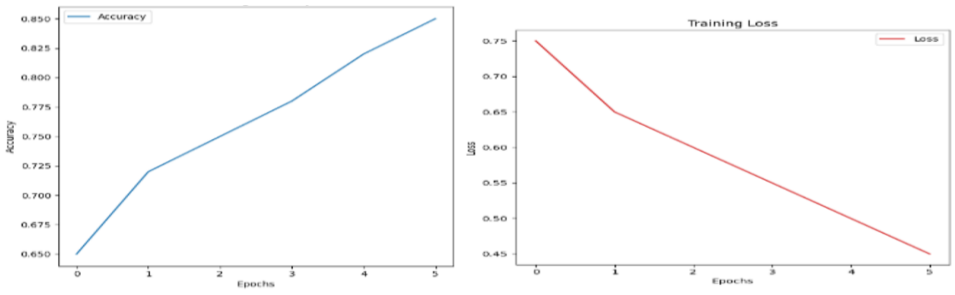
Table 6 compares the RCO-XLNet model to baseline models based on accuracy, runtime, and compute cost. The suggested RCO-XLNet model has the best accuracy (92.3) and the lowest runtime (6 hours) and compute cost (1000 units), which proves its efficiency. The CNN-LSTM, BERT, and Robustly Optimized BERT Pretraining Approach (RoBERTa) models perform well but are more expensive in terms of runtime and computation. Bayesian Optimization of hyperparameters is less accurate but also relatively cheap in terms of compute cost and runtime. This table shows the computational benefits of the RCO-XLNet model in practical use.

Table 6. Performance Comparison of RCO-XLNet with Baseline Models

Model	Accuracy (%)	Runtime (hours)	Compute Cost (Units)
RCO-XLNet [Proposed]	92.3	6	1000
CNN-LSTM [13]	88.0	8	1200
BERT [3]	89.5	10	1300
RoBERTa [16]	90.3	9	1250
Bayesian Optimization for Hyperparameters [1]	87.4	7	1100

4.6. Ablation and Sensitivity Analysis

Figure 9 shows the training curves for accuracy and loss over the course of training. The first subplot displays the accuracy curve, illustrating the steady improvement in accuracy as training progresses. The second subplot presents the loss curve, which decreases over time, indicating that the model is learning effectively. Both plots use data from the training process, showing the model's stability and convergence. These curves are essential for evaluating the model's performance during training.



a) Training Accuracy. b) Training loss.

Figure 9. Training accuracy and loss curves.

Figure 10 illustrates the sensitivity of the model's accuracy to different learning rates. The plot shows the accuracy of the model as a function of learning rate, ranging from 0.001 to 0.3. The curve demonstrates how performance improves with increasing learning rate up to a point before plateauing. This figure highlights the importance of selecting an optimal learning rate for model training.

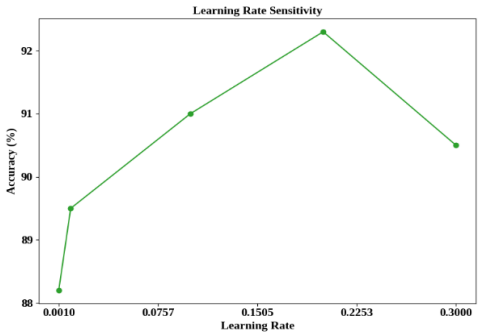


Figure 10. Learning rate sensitivity plot.

Table 7 presents the results of the ablation study, comparing different configurations of the RCO-XLNet model. The study evaluates performance using MFCC with only audio, XLNet with only lyrics, fusion of both audio and text, and fusion with RCO optimization. The table shows accuracy, precision, recall, and F1-score for each configuration. The results demonstrate the incremental value of RCO optimization, with the fusion + RCO model achieving the highest performance across all metrics.

Table 7. Ablation Study Results for RCO-XLNet Model.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RCO-XLNet (Proposed)	92.3	93.8	91.6	94.2
MFCC with only Audio	85.5	83.0	84.1	83.5
XLNet with only Lyrics	88.0	86.5	87.0	86.8
Fusion (Audio + Lyrics)	91.5	92.0	90.8	91.4
Fusion + RCO	94.2	95.1	93.5	94.8

5. Conclusions

The present study overcomes the drawbacks of conventional music emotion recognition methods by presenting a novel hybrid model called Redefined Crayfish Optimization-driven XLNet (RCO-XLNet). Using typical NLP algorithms and simple audio analysis, conventional approaches frequently suffer from issues with precision and lack of ability to grasp complex emotional expressions in music. The developed RCO-XLNet model is integrated with MFCC and TF-IDF feature extraction techniques in the suggested methodology to improve emotion classification. Compared to existing techniques, our method performs much better with 92.3% accuracy, 93.8% precision, 91.6% recall, and 94.2% F1-score. The key weaknesses of the RCO-XLNet model are extremely high computational resource requirements, sensitivity to noisy input data, possible biasing by the limited availability of lyrics, and lower generalizability to different genres, cultures, and languages. Also, it might be an issue of fairness with gendered or explicit content in lyrics, which can affect user experience and ethical usage. Confidence thresholds, human-in-the-loop validation, and on-device computation are some of the strategies that can be used to reduce these challenges. The deployment scenarios may involve therapeutic or mood-enhancing playlists, customized suggestions in streaming applications, and real-time emotion-sensitive music adaptation on mobile or wearable devices. Future research may explore its application to a wide range of musical styles, its combination with real-time music recommendation systems, and its expansion to multimodal emotion analysis. Besides, future studies may investigate contrastive audio-text training to enhance correspondence between lyrics and audio characteristics, scalable self-supervised learning to take advantage of large unlabeled music datasets, and more refined emotion representations using dimensional models like valence-arousal. Discrete emotion labels and continuous scales would also be useful in improving the interpretability and strength of music emotion recognition models.

Declarations

Funding: Authors did not receive any funding.

Conflicts of interests: Authors do not have any conflicts.

Data Availability Statement: No datasets were generated or analyzed during the current study.

Code availability: Not applicable.

Authors' Contributions

Ping Ran is responsible for designing the framework, analyzing the performance, validating the results, and writing the article. Kwen Liew is responsible for collecting the information required for the framework, provision of software, critical review, and administering the process.

References

[1] Bergstra J., Bardenet R., Bengio Y., and Kégl B., "Algorithms for Hyper-Parameter Optimization," *Advances in Neural Information Processing Systems*, vol. 24, pp. 2546-2554, 2011. https://papers.nips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf

[2] Chen C. and Li Q., "A Multimodal Music Emotion Classification Method Based on Multifeature Combined Network Classifier," *Mathematical Problems in Engineering*, vol. 2020, pp. 1-11, 2020. shttps://doi.org/10.1155/2020/4606027

[3] Cortiz D., "Exploring Transformers in Emotion Recognition: A Comparison of BERT, DistilBERT, RoBERTa, XLNet and ELECTRA," *arXiv preprint*, vol. arXiv:2104.02041, pp. 1-7, 2021. https://doi.org/10.48550/arXiv.2104.02041

[4] Deng S., Wu L., Shi G., Xing L., and et al, "Learning to Compose Diversified Prompts for Image Emotion Classification," *Computational Visual Media,* vol. 10, no. 6, pp. 1169-1183, 2024. https://doi.org/10.1007/s41095-023-0389-6

[5] Deng Y. and Lin N., "Analysis and Expression of Music Emotion Based on CAD and Deep Reinforcement Learning Algorithm," *Computer Aided Design and Applications Journal*, vol. 21, no. 23, pp. 19-34, 2024. doi:10.14733/cadaps.2024.S23.19-34

[6] Federico S., Francesca C., and Stavros N., "Joint Learning of Emotions in Music and Generalized Sounds," *arXiv preprint*, vol. arXiv:2408.02009, pp. 1-6, 2024. doi:10.48550/arXiv.2408.02009

[7] Fong H., Kumar V., and Sudhir K., "A Theory-Based Explainable Deep Learning Architecture for Music Emotion," *arXiv preprint*, vol. arXiv:2408.07113, pp. 1-54, 2024.

DOI:10.48550/arXiv.2408.07113

[8] Gomez-Cañón J., Cano E., Eerola T., Herrera P., and et al, "Music Emotion Recognition: Toward New, Robust Standards in Personalized and Context-Sensitive Applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106-114, 2021. DOI:10.1109/MSP.2021.3106232

[9] Grekow J., "Music Emotion Recognition Using Recurrent Neural Networks and Pretrained Models," *Journal of Intelligent Information Systems*, vol. 57, no. 3, pp. 531-546, 2021. https://doi.org/10.1007/s10844-021-00658-5

[10] Gudivaka B., "Designing AI-Assisted Music Teaching with Big Data Analysis," *Current Science and Humanities*, vol. 9, no. 4, pp. 1-14, 2021. DOI:10.5281/zenodo.12605388

[11] Gujar S. and Reha A., "Enhancing Accuracy and Performance in Music Mood Classification Through Fine-Tuned Machine Learning Methods," *Commun. Appl. Nonlinear Anal.*, vol. 31, no. 5s, pp. 234-258, 2024. doi:10.52783/cana.v31.1019

[12] He N., Neural Networks for Music Emotion Recognition and Social Tags Emotion Representation, Doctoral dissertation, 2023. http://hdl.handle.net/10453/171926

[13] Hizlisoy S., Yildirim S., and Tufekci Z., "Music Emotion Recognition Using Convolutional Long Short Term Memory Deep Neural Networks," *Engineering Science and Technology, an International Journal*, vol. 24, no. 3, pp. 760-767, 2021. https://doi.org/10.1016/j.jestch.2020.10.009

[14] Kamau N., *Emotion-Driven Music Recommendations: Integrating CNN and KNN for Personalized Playlists*, Doctoral dissertation, Dublin Business School, 2023. https://hdl.handle.net/10788/4400

[15] Lakshmana-Kumar R., Jayanthi S., Muthua B., and Sivaparthipan C., "An Automatic Anomaly Application Detection System in Mobile Devices Using FL-HTR-DBN and SKLD-SED K means algorithms," *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, vol. 46, no. 2, pp. 3245-3258, 2024. https://doi.org/10.3233/JIFS-233361

[16] Liu Y., Ott M., Goyal N., Du J., and et al, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint*, vol. arXiv:1907.11692, pp. 1-13, 2019. https://doi.org/10.48550/arXiv.1907.11692

[17] Liu Z., Xu W., Zhang W., and Jiang Q., "An Emotion-Based Personalized Music Recommendation Framework for Emotion Improvement," *Information Processing and Management.*, vol. 60, no. 3, pp. 103256, 2023. doi:10.1016/j.ipm.2022.103256

[18] Liyanarachchi R., Joshi A., and Meijering E., "A Survey on Multimodal Music Emotion Recognition," *arXiv preprint*, vol. arXiv:2504.18799, pp. 1-26, 2025. https://doi.org/10.48550/arXiv.2504.18799

[19] Medina Y., Beltrán J., and Baldassarri S., "Emotional Classification of Music Using Neural Networks with The Mediaeval Dataset," *Personal and Ubiquitous Computing*, vol. 26, no. 4, pp. 1237-1249, 2022. https://doi.org/10.1007/s00779-020-01393-4

[20] Moscato V., Picariello A., and Sperli G., "An Emotional Recommender System for Music," *IEEE Intelligent Systems*, vol. 36, no. 5, pp. 57-68, 2020. DOI:10.1109/MIS.2020.3026000

[21] Naser D. and Saha G., "Influence of Music Liking on EEG Based Emotion Recognition," *Biomedical Signal Processing and Control*, vol. 64, pp. 102251, 2021. https://doi.org/10.1016/j.bspc.2020.102251

[22] Panda R., Malheiro R., Rocha B., Oliveira A., and Paiva R., "Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis," *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research*, Marseille, pp. 1-13, 2013. https://api.semanticscholar.org/CorpusID:163157292

[23] Pandeya Y. and Lee J., "Deep Learning-Based Late Fusion of Multimodal Information for Emotion Classification of Music Video," *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2887-2905, 2021. https://doi.org/10.1007/s11042-020-08836-3

[24] Pandeya Y., Bhattarai B., and Lee J., "Deep-Learning-Based Multimodal Emotion Classification for Music Videos," *Sensors*, vol. 21, no. 14, pp. 4927, 2021. https://doi.org/10.3390/s21144927

[25] Rajesh S. and Nalini N., "Musical Instrument Emotion Recognition Using Deep Recurrent Neural Network," *Procedia Computer Science*, vol. 167, pp. 16-25, 2020, https://doi.org/10.1016/j.procs.2020.03.178

[26] Sajid S., Javed A., and Irtaza A., "An Effective Framework for Speech and Music Segregation," *The International Arab Journal of Information Technology*, vol. 17, no. 4, pp. 75-82, 2020. DOI:10.34028/iajit/17/4/9

[27] Sams A. and Zahra A., "Multimodal Music Emotion Recognition in Indonesian Songs Based on CNN-LSTM, XLNET Transformers," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 1, pp. 355-364, 2023. DOI:10.11591/eei.v12i1.4231

[28] Sarkar R., Choudhury S., Dutta S., Roy A., and Saha S., "Recognition of Emotion in Music Based on Deep Convolutional Neural Network," *Multimedia Tools and Applications*, vol. 79, no. 1, pp. 765-783, 2020.

https://doi.org/10.1007/s11042-019-08192-x
[29] Sheykhivand S., Mousavi Z., Rezaii T., and Farzamnia A., "Recognizing Emotions Evoked by Music Using CNN-LSTM Networks on EEG Signals," *IEEE Access*, vol. 8, pp. 139332-139345, 2020. DOI:10.1109/ACCESS.2020.3011882.
[30] Shi G., Deng S., Wang B., Feng C., and et al, "One for All: A Unified Generative Framework for Image Emotion Classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7057-7068, 2024. DOI:10.1109/TCSVT.2023.3341840
[31] Sujeesha A., Mala J., and Rajan R., "Automatic Music Mood Classification Using Multi-Modal Attention Framework," *Engineering Applications of Artificial Intelligence*, vol. 128, pp. 107355, 2024. DOI:10.1016/j.engappai.2023.107355.
[32] Wang S., Xu C., Ding A., and Tang Z., "A Novel Emotion-Aware Hybrid Music Recommendation Method Using Deep Neural Network," *Electronics*, vol. 10, no. 15, pp. 1769, 2021. https://doi.org/10.3390/electronics10151769
[33] Wang Y., "Music Composition and Emotion Recognition Using Big Data Technology and Neural Network Algorithm," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1-11, 2021. https://doi.org/10.1155/2021/5398922
[34] Wu J., "Research on Music Emotion Analysis and Dance Creation Based on Neural Network," *Journal of Electrical Systems*, vol. 20, no. 6s, pp. 575-581, 2024.
[35] Zainab R. and Majid M., "Emotion Recognition Based on EEG Signals in Response to Bilingual Music Tracks," *The International Arab Journal of Information* Technology, vol. 18, no. 3, pp. 26-36, 2021, DOI:10.34028/iajit/18/3/4
[36] Zhang S., Huang Z., and Lang Y., "Application of Video Game Algorithm Based On Deep Q-Network Learning in Music Rhythm Teaching," *The International Arab Journal of Information Technology*, vol. 22, no. 1, pp. 124-138, 2025. DOI:10.34028/iajit/22/1/10

Ping Ran was born in SIchuan, Xuanhan. China, in 1978. received her master's degree from Shaanxi Normal University in 2007. Now she is the Dean and Associate Professor of Guqin Department in Xi'an International University. Her research interests include Vocal Music Teaching and Performance, Qin Song Teaching and Performance, and Musical Intelligences.