Minimax Theory and its Applications

Volume 11 (2026), No. 2s, 544 to 559

Hybrid soft computing technique for music emotion recognition based on deep feature extraction and optimal feature selection with F-CNN+ELNN

Dr. R. Sundar[1], Kavibharath L[2], Dr. R. Hema[3], G.Manikandan[4], Dr. M. Balakrishnan [5], Dr.N.Senthil Madasamy[6], Dr. Revathy P[7]

[1]Associate Professor, Department of Marine Engineering, AMET Deemed to be University 135, ECR Road, Kanathur, Chennai, Tamil Nadu, 603112, India. Email: sundar.r@ametuniv.ac.in,ORCID ID:0000-0002-7969-410X

[2]Assistant Professor , KalaignarKarunanidhi Institute of Technology, Coimbatore. Email: lkavibharath06@gmail.com, Orcid: 0009-0001-5588-4013

[3]Assistant Professor ,Department of Computer Science SDNB Vaishnav College for Women ,Chromepet,Chennai-44. Email: hemaramji@sdnbvc.edu.in

[4]Professor, Department of Artificial Intelligence and Machine Learning, St.Joseph's College of Engineering, Chennai-119, India. Email: maniprofessional@gmail.com,ORCID ID:0000-0002-1184-8387

[5]Professor, Department of Artificial Intelligence and Data Science, Dr. Mahalingam College of Engineering and Technology, Pollachi, Email: balakrishnanme@gmail.com

[6]Associate Professor of Computer Science, Dr.Mahalingam College of Engineering and Technology Udumalai Road, Pollachi- 03. Email: senkav1293@gmail.com, ORCID ID:0000-0003-4486-3281

[7]Assistant Professor , Department of Mathematics Sri Krishna College of Engineering and Technology, Coimbatore – 8, Tamil Nadu, India.Email: revaprabhu92@gmail.com, ORCID ID: 0000-0001-8531-4700

Abstract

Music holds a significant place in human history, serving as a means to convey emotions. With the growing importance of music emotion recognition in digital entertainment and human-machine interaction, there has been considerable advancement in this field. However, existing research often relies on short-term artificial stimuli, limiting their ability to evoke complex emotions and capture dynamic brain responses. While previous studies have segmented music comprehensively at similar intervals to analyze mood, recent efforts focus on recognizing mood by extracting sections with similar acoustic features. In our study, we propose a hybrid soft computing approach for music emotion recognition, integrating deep feature extraction and optimal feature selection techniques. Initially, we employ a feeding convolutional neural network (F-CNN) layer to extract deep features from audio signals. Subsequently, we introduce a modified wolf pack (MWP) cluster-based algorithm to optimize feature clustering, addressing data dimensionality challenges effectively. Additionally, we utilize an enhanced Laguerre neural network (ELNN) model for predicting emotional dimensions' levels. Our proposed method is evaluated using the Soundtrack dataset comprising 360 sound samples. The F-CNN+ELNN model achieves an impressive maximum accuracy of 97.568%, outperforming existing models by 15.689%. It signifies a significant advancement in music emotion recognition, demonstrating the efficacy of our approach in accurately capturing and interpreting emotional nuances in music.

Keywords: music emotion, soft computing, cluster-driven, deep feature extraction, optimal feature selection

1. Introduction

Today's popular music is incredibly diverse and rich in aspects. Due to issues with intellectual property, the production process requires a large number of experts and takes a long period. The cost of authorization or cooperation both demand large sums of money. In current years, computer-automated composition has steadily increased prominence as a solution to the aforesaid subjects. Pre-calculated exact replicas or machine learning schemes are typically applied as aid to produce the computer-automated composing. These systems use standardized algorithms to produce random music based on related fundamental music ideas including pitch, rhythm, and chords. The recording created by the Hidden Markov Model (HMM) is one of the more well-known

ones [1]. The neural network, sometimes referred to as the artificial neural network (ANN), has trusted heavily on hardware calculation due to the quick growth of knowledge and technology. The HMM requires extra model aspects in addition to its statistical foundation, which can significantly lessen the pre-work involved in creating music. This article employs computer automation that resembles a neural network to compose music, allowing users to easily create music that expresses their current feelings. Particularly when it comes to music, human emotions are incredibly complicated [2][3]. Different researchers define the fundamental categories of emotions differently. For instance, using facial expression analysis, Ekman identified the six basic emotions in 1972 [4]. In order to allocate shared feelings in a two-dimensional plane and discover more about the association between various emotions, Russell developed the emotional circle model in 1980 [5]. This model employs the straight axis and the positive and negative feelings as the upright axis. Researchers have used the two-dimensional plane model to analyze and quantify emotions in great detail. This has allowed them to investigate the connections between emotions in diverse domains. Gomez et al. investigated the connection among harmonious features and two-dimensional emotional planes in 2007 [6]. As similar neural networks have developed, a wide variety of models, including DNN, CNN, RNN, and GAN, consume emerged. Research on the subject of machine learning and melody has been conducted; two examples of these studies are MuseGAN [8] and Yang YixuanMidiNet [7]. Numerous monotonous tasks, such as MIR music information retrieval and music theory analysis, were required for the reduced HMM. More non-music specialists could now more readily enter the field of composition research thanks to the acceleration of automated composition's efficiency. However, many aspects that would have been taken into account in the early automated composition-emotions were overlooked by these new methods. These days, academics frequently just concentrate on two of the three fields—machine learning, emotion, and music—and talk about their relative importance.

Next, songs that match to different emotions are created, which acts as a basic prototype for multidisciplinary study and progressively streamline the song conversion process. The emotional circle model, which Russell created, serves as the foundation for emotion-related research. Since then, it has been expanded to include more areas and facets. For instance, Laurier et al. (2009) developed a novel emotional circulation plan, depicted in Fig. 1, by using self-organized mapping (SOM, Self-organizing Map), and a two-dimensional emotional plane as the foundation [9]. It is possible to observe similarities between the distribution of emotional terms and the group classification trend. Furthermore, Yang Yixuan, ZhengShikang, and others studied and explored the two kinds of emotion organization in melody using various calculation models [10].

For a very long time, human communities have required music. Both informal social gatherings and official state occasions feature it. In an effort to affect the emotions of those who are listening to it, movie soundtracks are employed to heighten suspense, create tension, and more. There is a widespread consensus regarding the emotion or emotions observed, even though not everyone reacts to a musical excerpt in the same way. Based on specific musical elements, [11] created an algorithm to predict emotion in music. Anger, fear, happiness, sadness, and tenderness were the five main emotional categories that were targeted. Twelve specialists in music chose the pieces, which totaled fifty extracts and featured both strong and mild representations of each of the five emotions. By employing a bigger, unscreened music information set and linking it to a broader variety of identifiable emotions, this study expands on the findings. The continuous aspect of emotions was overlooked when assessing dispositions using a set number of separate categories, according to Wood and Semwal [12]. Restricting study to a carefully specified subsection of music where a single emotion is extremely characterized misses possible side effects in algorithms when run on music having numerous, imbrication emotions, meanwhile a lot of music involves a mix of emotions. In [13] examined earlier studies that focused on categorizing distinct musical structural elements and the feelings they elicited. They observed that the same or comparable emotion assessments were connected to a number of the structural elements. For instance, activation is affected by tempo, loudness, and timbre.Around 400,000 new instances of cancer are reported in families worldwide each year, according to data from the World Health Organization (WHO) [14]. The rate of survival in industrialized nations is more than 80% because to advancements in diagnosis methods and medical care. However, because instances are significantly underdiagnosed, particularly in South Asia and sub-Saharan Africa [15], it is less than 30% in developing nations [16]. In addition to the intrusive testing and aggressive treatment that can cause them great physical pain, these patients may not fully understand what is going on around them, and they may be cut off from their friends, school, and daily life [17].The emotions, sentiments, and mental state are definitely affected by all of these changes, which lower their excellence of lifetime. Within the field of pediatric oncology, [18] music therapy (MT) is an internationally recognized interference that is recommended by the doctors [19]. Previous studies have shown that MT can have effect on children with terminal cancer's emotional components of treatment, such as worry and fear. In this sense, [20] scholars in the fields of MT, psychology, and neurology, among others, have been examining the connection between music and emotions for decades.

For further enhancement, we present the hybrid soft computing framework for music emotion recognition using deep feature extraction and optimal feature selection. The major charities of proposed work are abridged as follows.

1. The use of feeding convolutional neural network (F-CNN) facilitates the extraction of deep features from audio signals. F-CNN adept at capturing complex acoustic patterns and subtle emotional cues, providing rich representations of the music data.
2. A modified wolf pack (MWP) cluster based algorithm is used for the optimal feature clustering, effectively addressing challenges related to data dimensionality.
3. Our study incorporates an enhanced Laguerre neural network (ELNN) model used for predicting emotional dimensions' levels. ELNN model uses advanced neural network to provide accurate and nuanced interpretations of emotional nuances present in music.
4. We evaluate the proposed method using the Soundtrack dataset, which comprises 360 sound samples representing diverse emotional states. Through rigorous evaluation, we demonstrate the efficacy and superiority of our approach in accurately capturing and interpreting emotional nuances in melody.

The rest of this paper is prearranged as follows. In Section 2, we current a detailed appraisal of the prevailing literature concerning the Hybrid soft computing technique for music emotion recognition. Section 3 delves into the research methodologies and principles employed in this study, including the deep feature extraction and optimal feature selection. Section 4 conducts comparative analysis and provides comprehensive assessment of theemotion recognition usingfeature extraction, selection and outcomes using various metrics. In Section 5, the conclusions and expectations of this research are discussed.

2. Related works

2.1 State of the Art

Emotion in [21] music is crucial for listeners' cognitive processes. The speed at which technology is developing has led to a greater diversity and quicker dissemination of music. The price of producing music is still very high, though. In recent years, AI music composition has steadily attracted interest as a answer to the problem. The area of this investigate is to progress an automated system for composition that integrates machine learning, music, and emotions. The CVAE-GAN perfect, which serves as the architecture for the system, is trained through deep learning on the music database tagged with emotions. This allows the model to generate music segments that correlate to the designated emotions. After hearing the system's output, the participants determine that the music reflects their initial feelings.

Quantifying [22] the emotions elicited by music can be challenging. A person's subjective emotions are different from the concrete characteristic data of a certain music snippet. A being's awareness of music can be influenced by their own experiences, but the combination of musical elements—such as rhythm, timbre, and tone—all affects the sentiment that is felt. It is also challenging to quantify the emotions elicited by music using a limited sample of emotions since people frequently link a variety of emotions with a single piece of music. This study describes a series of algorithms to associate intricate combinations of emotion categories with detailed mouth data that was taken from a collection of instrumental music passages.

Algorithmic and automated [23] music arrangement are grounded on logic operations, where music limitations are set to reflect the envisioned mood or style. With the right charting techniques, computer-generated music can be combined with other domains, like intermediary arts and music technology. The foremost focus of this paper is on the possibilities of utilizing the emotion categorization and parameters to merge motion devices and automatic composition with an emotional documentation system. The emotion map allows for the simultaneous control of the corresponding motion path and music pattern by time-varying cursor movement. In line with this, an interactive music-motion platform is created.

The popularity of Chinese [24] is increasing as a result of the growth of Chinese language education. The few amounts of Chinese characters in Tang Dynasty poetry convey a great deal of information. As a result, it might be difficult to help students comprehend the emotional content and backstory of poetry. To enhance the learning of poetry appreciation by offering three multimedia properties to support immersive poetry learning: interpretation with sporadic poetry, emotional support with contextual music, and vision with scenic spots. An annotation the context material poem contents and music emotional qualities in order to constructs the immersive poetry learning with software setting. The learner contexts can serve as the foundation for the ontology-based immersive poetry learning organization, which uses ontology mapping to produce coherent poetry and music.

The WHO Global [25] Creativity on infantile cancer was established in 2018 with the goal of saving an extra million lives by 2030 by achieving at least a 60% existence rate for children with malignance while also minimizing their anguish. India has a chance to advance and connect itself with this global initiative. Several stakeholders have been involved in a number of meetings held at various platforms in India and outside during the past year. As a result, there is now a greater understanding of the Global Initiative, a consensus among Indians to participate in it, a list of important participants, and preliminary conversations on the areas of emphasis and pertinent organizations. They plan to outline the specifics of the project's scope, activity structure, timetables, intermediate and final goals, and required resources within the next year.

Over the past [26] few years, the study of teacher emotions has drawn more and more attention from researchers. Different tools to measure teachers' emotions have been developed, and quantitative methods have been added to the initial, primarily qualitative research methods. The goal of the current study was to validate the Teacher Emotions Scale (TES) in Kosovo, a low-to-middle-income nation in Southeast Europe. Results from N = 258 instructors in Kosovo show that TES-Albanian functions similarly to the German original and the English version of TES in terms of factor structure. These findings lend support to the external validity of the scale. They draw the conclusion that TES is a useful tool for measuring teacher emotions in non-WEIRD cultural contexts when translated into other languages.

In order to intelligently [27] analyzed how different music types affect children's emotions, this research builds a multimodal convolutional neural network perfect and conducts a thorough investigation of how different music genres affect children's emotions. The auditory power spectrograph gotten by feature withdrawal is cut using the MSCN-LSTM. This is done in consideration of the assortment of music type landscapes in the audio power spectrograph. The long and petite period memory network and the multi-scale convolutional kernel convolutional neural network make up the MSCN-LSTM. The chronological features of the eye-movement signal are extracted using the LSTM network, the features of the EEG signal are extracted using the MSCNN network, and feature-level fusion is performed. The multimodal signal for children achieves an average accuracy of quadruple categorization, which is 97.94%.

In order to determine the most significant EEG characteristics triggered by music expressing various emotions [28] and carried out number of tests. From each of the 12 electrode sites, we extracted 27-dimensional topographies. They utilized a correlation-based feature assortment technique to find the feature set that had the lowest redundancy and the strongest relationship to the original features. The number of classifiers was employed to assess the innovative and selected feature sets' recognition accuracy, including SVM, C4.5, LDA, and BPNN. Ultimately, a thorough analysis of the data is presented, demonstrating the evident connections between the chosen feature set and human emotions.

A transformer-based multi-modal convolutional neural network [29] is used for the purpose of determining how music genres affect children's emotions. The filmed and acoustic features are extracted using the BiLSTM sub-network perfect, and the text topographies are extracted using the BERT sub-network. They successfully fuse various forms of modal information by utilizing the enhanced transformer cross-modal fusion module. The transformer module evaluates data from many modalities and extracts emotion from the multimodal data. The numerous studies shows that the multi-modal CNN model and outperforms previous approaches in footings of prediction correctness and significantly raises the accuracy of sentimentality classification tasks.

A personalized recommendation system [30] was made based on the interests and purchasing habits of the users. When it comes to handling unclear information, evidence theory excels in differentiating between true and erroneous information. In order to do this, the evidence theory algorithm in this article's study of the application of the recommendation system can infer the information users require based on the ambiguous data. Furthermore, the experiment described in this article demonstrates how well an algorithm applying evidence theory to a recommendation system can understand user preferences and suggest information that consumers require.

2.2 Problem description

Han et al. [31] proposed neural network architecture for music emotion recognition, utilizing an Inception-GRU residual structure. In our investigation of music emotion gratitude, we grapple with the intricate task of accurately discerning and categorizing the emotional nuances present in music signals. Music's complexity lies in its ability to evoke a diverse range of emotions, from joy and excitement to sadness and nostalgia, often simultaneously or in rapid succession [21]. Conventional emotion recognition models often struggle to capture these nuances effectively, given the high-dimensional and heterogeneous nature of music data [22]-[23]. The intricate interplay of auditory topographies such as pitch, rhythm, tempo, timbre, and dynamics necessitates a

sophisticated approach to feature extraction and modeling capable of discerning subtle variations and patterns indicative of different emotional states [24][25]. Annotated datasets for music emotion recognition are frequently limited in size and diversity, posing challenges for effectively training ML/DL models. The scarcity of labeled data [26] can hinder models' generalization capability, potentially leading to overfitting or biased outcomes. Extracting informative features from raw audio signals is a crucial aspect of music emotion recognition [27]. Traditional feature extraction methods may overlook subtle emotional cues present in the audio, thereby limiting model performance [21]-[30]. DL-based feature extraction offer promise in addressing this challenge but often necessitate considerable quantities of labeled data for effective exercise [28]. DL models for music emotion recognition computationally demanding, necessitating significant computational resources for training and inference [29][30]. The complexity may constrain the scalability and real-time use of ML/DL-based approaches, particularly in resource-constrained environments. To address these challenges, we proposed a hybrid soft computing approach that combines deep feature extraction and optimal feature selection techniques tailored specifically for music emotion recognition. The proposed work consists of following research objectives:

1. Design and implement advanced feature extraction techniques tailored specifically for music emotion recognition, capable of capturing subtle emotional cues present in raw audio signals.
2. Collaborate with experts in musicology and psychology to ensure comprehensive coverage of emotional states across different musical genres and cultural contexts.
3. Develop efficient DL architectures and training algorithms used for music emotion recognition tasks, optimizing computational resources without sacrificing model performance. It reduces the computational complexity and enables real-time inference on resource-constrained devices.
4. Incorporate mechanisms for optimization model for predictions and understanding the underlying emotional cues encoded in music data.

3. Proposed methodology

In this segment, we define the working process of a hybrid soft computing approach for music emotion recognition such as deep feature extraction and optimal feature selection techniques. The hybrid soft computing approach for music emotion recognition, as depicted in Fig. 1, begins with the input of soundtrack dataset containing audio samples labeled with emotional categories such as sad, anger, tender, and happy. The signal preprocessing techniques may be applied to improve the excellence of the audio indications or remove sound. Subsequently, spatio-temporal features are extracted from the preprocessed audio signals using an F-CNN, specialized for sequential data like audio. This process enables the extraction of intricate patterns and variations in the audio signals indicative of different emotional states. To optimize the clustering of the extracted features, a modified wolf pack (MWP) clustering algorithm is used, effectively reducing data dimensionality while preserving relevant discriminative information. The optimized feature clusters obtained from MWP serve as input to the enhanced Laguerre neural network (ELNN) model, which is specifically designed for processing time-series data with nonlinear dynamics, making it well-suited for modeling the temporal nuances present in music. In inference, the trained ELNN model accurately predicts the emotional category of new music samples based on their extracted features, thereby enabling precise music emotion recognition.

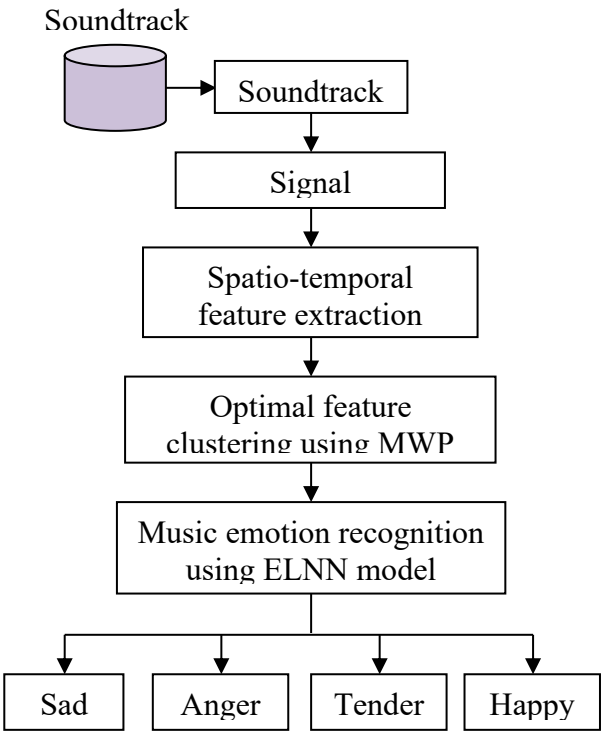


Fig. 1 Hybrid soft computing approach for music emotion recognition

3.1 Deep feature extraction using F-CNN

The process of deep feature extraction using feeding convolutional neural network (F-CNN) involves several technical steps aimed at uncovering meaningful insights from the input music signals. The architecture comprises multiple convolutional layers designed to extract hierarchical features from the spectrograms of the audio signals. Through convolutional filters and pooling operations, the network identifies local patterns and spatial dependencies, enhancing its capability to discern essential features within the music data. It enables the F-CNN to extract abstract and high-level representations of emotional content embedded within the music signals. As the data propagate through the network, feature maps are generated, with each map capturing specific patterns and characteristics present in the music data. Through representation learning, the F-CNN autonomously learns to extract discriminative features relevant for music emotion recognition tasks. The process of deep feature extraction using F-CNN involves the following working process.

1. F-CNN is meticulously crafted to extract hierarchical features from spectrograms of audio signals. It plays a crucial role in capturing intricate acoustic patterns and subtle emotional cues present in the music data.
2. Spectrograms, which provide a detailed representation of the frequency content of audio signals over time, serve as the input to the F-CNN. These spectrograms are generated using signal processing techniques and represent essential acoustic features of the music data.
3. During training, the network extract discriminative features from the spectrograms that are indicative of different emotional states or characteristics present in the music.
4. The trained F-CNN is used to extract deep features from new or unseen audio samples. These deep features represent the underlying acoustic patterns and emotional cues present in the music data and serve as input to downstream tasks such as emotion recognition or classification.
5. To prevent overfitting and improve generalization capability, dimensionality reduction such as fully connected layers and dropout layers used. It helps reduce the dimensionality of the extracted features while preserving essential information, enhancing the robustness of the emotion recognition system.

3.2 Feature clustering using MWP algorithm

Feature clustering using the modified wolf pack (MWP) algorithm represents a pivotal step in our approach to music emotion recognition. The MWP algorithm, inspired by the social behavior of wolf packs, offers a sophisticated mechanism for optimizing feature clustering, thereby addressing challenges related to data dimensionality and enhancing the discriminative power of the extracted features. The process of feature clustering involves grouping similar features together based on their characteristics, allowing the extraction of relevant patterns and constructions from the music data. It operates by iteratively optimizing the clustering of

features using a combination of exploration and exploitation strategies. Through the interaction and collaboration among the wolves, the algorithm effectively partitions the feature space into clusters that encapsulate distinct patterns and characteristics of the music data. Respectively gray wolf is the prey and compute the objective function is compute as follows.

$$\vec{C} = \left| \vec{D} \cdot \vec{P}_X(t) - \vec{P}(t) \right| \quad (1)$$

$$\vec{P}(t+1) = \vec{P}_X(t) - \vec{A} \cdot \vec{C} \quad (2)$$

where t is the current iteration $\vec{P}_X(t)$ denotes the position of the predator and $\vec{P}(t+1)$ denotes the position of the gray wolves in the next repetition. $\vec{A}$ and $\vec{C}$ are coefficient variable star and are articulated as $\vec{A} = 2a \cdot r_1 - a$ and $\vec{C} = 2.r_2$, where a is on the grid [2,0] and r_1 and r_2 are random numbers. The first 3 solutions were adopted to device the social order (α, β, δ). The mathematical reckonings for this supposition are imagined as follows.

$$\vec{C}_\alpha = \left| \vec{D}_1 \cdot \vec{P}_\alpha - \vec{P} \right| \vec{C}_P = \left| \vec{C}_2 \cdot \vec{P}_X - \vec{P} \right| \vec{D}_\delta = \left| \vec{C}_3 \cdot \vec{P}_\delta - \vec{P} \right| \quad (3)$$

$$\vec{P}_1 = \vec{P}_\alpha - \vec{A}_1 \cdot \left(\vec{C}_\alpha \right) \vec{P}_2 = \vec{P}_\beta - \vec{A}_2 \cdot \left(\vec{C}_\beta \right) \vec{P}_3 = \vec{P}_\delta - \vec{A}_3 \cdot \left(\vec{C}_\delta \right) \quad (4)$$

$$\vec{P}(t+1) = \frac{\vec{P}_1 + \vec{P}_2 + \vec{P}_3}{3} \quad (5)$$

Accordingly, standards of 0 or 1 reproduce two dissimilar levels of band assortment, and the transfer function is intended as follows.

$$\vec{R}(t) = \begin{cases} 1 \; if \; rand \le \left| \dfrac{1}{1 + \exp\left(-\overline{P}(t)/\tau \right)} \right| \\ 0 \; otherwise \end{cases} \quad (6)$$

As a outcome, it is accepted to enterprise the classical of MGWO, which is articulated as follows.

$$\vec{D}(t+1) = \left| \sin\left(\frac{a}{\vec{D}(t)} \right) \right| \quad (7)$$

To improve the optimization aptitude of GWO, the ICMIC operator enables directory correction among gray Casanovas, and the parameter τ of the transfer function is perturbed by the perturbation function.

$$\vec{P}(t+1) = \omega_1 \cdot \vec{P}_1 + \omega_2 \cdot \vec{P}_2 + (1 - \omega_1 - \omega_2) \cdot \vec{P}_3 \quad (8)$$

The gradient magnitude (GM) and relative gradient orientation (RO) are individually assumed as follow.

$$\angle \nabla \mathrm{I}(x,y)\mathrm{I}GM = \sqrt{\mathrm{I}_X^2(i,j)+\mathrm{I}_y^2(i,j)} \tag{9}$$

$$\angle \nabla \mathrm{I}(x,y)_{RO} = \angle \nabla \mathrm{I}(x,y) - \angle \Delta \mathrm{I}(x,y)_A \tag{10}$$

$$\angle\nabla \mathrm{I}(x,y) = \arctan\left|\frac{\left(\mathrm{I}_y(i,j)\right)}{\left(\mathrm{I}_x(i,j)\right)}\right| \quad (11)$$

$$\angle\nabla \mathrm{I}(x,y)_A = arcton\left|\frac{\left(\mathrm{I}_y(i,j)_A\right)}{\left(\mathrm{I}_x(i,j)_A\right)}\right| \quad (12)$$

$$\angle\nabla \mathrm{I}(x,y)\mathrm{I}GM = \sqrt{\mathrm{I}_X^2(i,j)+\mathrm{I}_y^2(i,j)} \quad (13)$$

Thus, a 3D feature vector is applied to estimation the excellence of bands:

$$Feature = \left[\upsilon \quad \upsilon \quad \upsilon \quad \right]_{GM,\ RO,\ RM} \quad (14)$$

Therefore, the appraisal principle is appeared as below

$$F(i) = \lambda \cdot OA(i) + (1-\lambda).\log_{10}\frac{n_c}{n_s(i)} \quad (15)$$

The feature clustering using MWP algorithm theatres a main role in the music emotion recognition process by organizing extracted features into cohesive groups, thereby simplifying subsequent analysis and interpretation.

3.3 Music emotion prediction using ELNN model

The soft computing techniques are used to predict emotional dimensions' levels based on the deep features extracted from audio signals using advanced neural network architectures. The process begins with the preparation of labeled music datasets, where each music sample is associated with emotional labels indicating its emotional content. These labeled datasets serve as the training data for the enhanced Laguerre neural network (ELNN) model, enabling it to learn the complex relationships between audio features and emotional dimensions. The ELNN model architecture consists of multiple layers of neurons, including input, hidden, and output layers, organized in a hierarchical fashion to capture the intricate patterns and structures present in the music data. The model incorporates Laguerre functions, a class of orthogonal functions known for their ability to model complex dynamic systems, as activation functions in the hidden layers. This choice of activation functions allows the ELNN model to capture nonlinear relationships between input features and emotional dimensions, enhancing its capability to predict nuanced emotional responses to music stimuli. During the training phase, the ELNN model iteratively adjusts its parameters using optimization techniques such as gradient descent and back propagation to minimize prediction errors and maximize the accuracy of emotional dimension predictions. The model learns to map the extracted deep features from audio signals to the corresponding emotional dimensions' levels, effectively capturing the subtle variations and nuances in emotional content present in the music data. Once trained, the ELNN model can be used to predict emotional dimensions' levels for new or unseen music samples, providing valuable insights into the emotional responses elicited by different music stimuli. When modeling a typical dynamic system using an ELNN model, time-delayed forms of contributions and productions are fed into the ELNN model. In general, a linear discrete-time system is represented by a nonlinear charting of past outputs and inputs.

$$b(j) = F(U(j),U(j-1),...,U(j-p),b(j-1),...,b(j-q)) \quad (16)$$

where F (·) is a nonlinear operative like a neural network or graph ELNN model, where m and n are the amount of forward rapports for input (U) and output (b), correspondingly. If the nonlinear scheme has limited recollection, i.e., p = , which implies that past output conditions are not reproduced, its discrete-time model is calculated as follows.

$$b(j) = F(U(j),U(j-1),...,U(j-p)) \quad (17)$$

where m becomes the computer's memory and the input is first processed by a standard linear time-invariant system and fed to a non-linear mapping block that produces the output. Laguerre bases are the set of orthonormal functions containing the interval of quadratic integral functions on the interval (0, ∞). The roots of

the logarithm are common in both the time and frequency domains. Orthonormality in the time domain is defined as follows.

$$\langle F,h\rangle = \int_0^{\infty} F(s)h(s)z(s)\ Ds = \begin{cases} 0\ if\ F \neq h \\ 1\ if\ F = h \end{cases} \quad (18)$$

where $z(s)$ is the allowance purpose. Otherwise, if $z(s)$ = 1, using Parseval's function, perpendicularity can be distinct in the incidence province as follows.

$$\langle F,h\rangle = \frac{1}{\overline{2\pi}} \int_{-\infty}^{\infty} F(jz)H(kz)\ Dz = \begin{cases} 0\ if\ F \neq H \\ 1\ if\ F = H \end{cases} \quad (19)$$

where f and H are Fourier converts of F and h, correspondingly. The Li is the desire answer of the ith Laguerre strainer, r is the amount of Laguerre sieves used, and ∗ means time sphere intricacy.

$$b(j) = F(L_0(j)*U(j), l_0(j)*U(j), ...L_R(j)*U(j), b(j-1)) \quad (20)$$

The chronological description of disconnected time Laguerre foundation purpose of instruction R and directory j is given as follows.

$$L_R[j] = x^{(j-R)/2} \sqrt{1-x} \sum_{k=0}^{R} (-1)^{k-R} \binom{j}{k} \binom{R}{k} b^{(R-k)} (1-x)^k \quad (21)$$

where 0 <x< 1 is the extremity stricture that regulates the rate of exponential asymptotic decline of the meanings. The z-transform of these foundation meanings is given as follows.

$$l_R(w) = \frac{\sqrt{1-x^2}}{w-x} \left(\frac{1-xw}{w-x} \right)^{R-1} \quad (22)$$

The extremity stricture controls how fast the instinct replies of lager strainers decay. When the pole stricture a is 0, the logger strainers developed fixed-delay operatives, as we can see from their z-transform. The rule function for ELNN model is defined as follows.

$$r_k : if\ a_1\ is\ X_1^k\ and\ a_2\ is\ X_2^k\ and ... and\ a_q\ is\ X_q^k,\ then\ b^k = H_k(a) \quad (23)$$

The fitness function can be updated by using the following function.

$$F(a) = \frac{\sum_{k=1}^{P} H_k(a) \prod_{i=1}^{q} \mu X_I^k(a_I)}{\sum_{k=1}^{P} \prod_{i=1}^{q} \mu X_I^k(a^I)} \quad (24)$$

where k= 1, 2, ... , M and I= 1, 2, ... , n are the instruction and contribution catalogues, correspondingly, and $\mu X_I^k(a_I)$ are the corresponding Gaussian membership functions (GMF).

$$\mu X_I^k(a_I) = Exp(-((a_I^k - p_I^k)/\sigma_I^k)^2) \quad (25)$$

where p_I^k and σ_I^k are the middles and scaling strictures of the GMF correspondingly. The mean squared error (MSE) is designated now as the error meaning, calculate as follows.

$$e(m) = \frac{1}{Q}\sum_{k=1}^{Q}[b(m, j) - b_D(j)]^2 \quad (26)$$

where $b_D(j)$ is the production of the actual arrangement or wanted production, $b(m, j)$ is the output of the perfect which will be recognized, P is the amount of drill data opinions, and j is the time directory. The limitation course m is entrenched in $b(m, j)$. For ease, $b(m, j)$ and $b_D(j)$ are referenced as b and b_D respectively.

4. Results and Discussion

In this segment, we present the outcomes and comparative analysis of proposed F-CNN+ELNN model for music emotion recognition, alongside existing models. To evaluate the effectiveness of our model, we employ the Soundtrack dataset comprising 360 sound samples. The implementation of our model is approved out using the Python programming language on the Google Colaboratory platform, leveraging libraries such as TensorFlow, Keras, scikit-learn, Matplotlib, and Seaborn. We benchmark the performance of our F-CNN+ELNN model against several existing models, including SVM + ReliefF, VGG16, Inception-v3, ResNet-18, DCNN [32], CNN [33] and CNN+Inception-GRU [31]. Quality metrics such as accuracy, precision, recall, and F1-score are utilized to measure and compare the efficacy of different music emotion recognition models.

Table 1. Dataset description

Class	Number of sound samples		
	Training	Testing	Total
Sad	60	30	90
Anger	60	30	90
Tender	60	30	90
Happy	60	30	90
Total	240	120	360

4.1 Dataset description

Founded on audio feature dispensation and deep learning, an experiment was conducted on a database of audio recordings. Four emotions were found in the test. The four emotions are happiness, anger, sadness and tenderness correspondingly. The recording dataset consists of 360 sound examples, music extracted from movie soundtracks every 30 seconds. Image feature recognition based on spectral features is used to combine the phantom topographies of the sensory info area in audio data to categorize music datasets. The dataset is based on the emotional arousal/valence 2D emotion perfect database. The arousal axis is scored from 1 to 8 from mild/sleepy to tense/excited, and the valency axis is scored from 1 to 8 from sad/frustrated to happy/happy. Sensory categorization of data comes from the auditory perception of the human body, not analyzing its tone from a reference base. However, it is important to note that the audio data is derived from film music with notable emotional characteristics. A musical resonance is an instrumental presentation that has no human voice or textual characteristics, and is stored in audio mp3 stereo format at 44.1 kHz. In the experiment, 156 audio clips were rated as angry, 58 happy, 68 sad, and 78 light. The ratio of exercise set to test set is 8:2. The test is completed by plotting the cross-entropy among the predictor and the loss function. The optimization technique used was Adam and the knowledge rate was set to 1e-3. For each class, the table specifies the number of sound samples allocated for training and testing purposes. In total, the dataset comprises 360 sound samples, with 240 samples allocated for training and 120 samples for testing across all emotional classes. This balanced distribution ensures sufficient data representation for each emotional category, facilitating robust model training and evaluation.

Table 2. Class wise results of F-CNN+ELNN model for music emotion recognition

Class	Epochs				
	20	40	60	80	100
	Accuracy (%)				
Sad	95.633	95.678	95.723	95.768	95.813
Anger	94.856	94.901	94.946	94.991	95.036
Tender	97.125	97.170	97.215	97.260	97.305
Happy	96.989	97.034	97.079	97.124	97.169
	Precision (%)				
Sad	94.856	94.901	94.946	94.991	95.036

Anger	94.353	94.398	94.443	94.488	94.533
Tender	94.867	94.912	94.957	95.002	95.047
Happy	94.633	94.678	94.723	94.768	94.813
	Recall (%)				
Sad	93.969	94.014	94.059	94.104	94.149
Anger	94.526	94.571	94.616	94.661	94.706
Tender	93.895	93.940	93.985	94.030	94.075
Happy	95.898	95.943	95.988	96.033	96.078
	F1-score (%)				
Sad	94.410	94.455	94.500	94.545	94.590
Anger	94.439	94.484	94.529	94.574	94.619
Tender	94.378	94.423	94.468	94.514	94.559
Happy	95.261	95.306	95.351	95.396	95.441

4.2 Comparative analysis

Table 2 presents the class-wise results of the F-CNN+ELNN model for music emotion recognition across different epochs, ranging from 20 to 100. The results are evaluated based on accuracy, precision, recall, and F1-score metrics for each emotion class: Sad, Anger, Tender, and Happy. Beginning with the accuracy metric, we observe consistent increases across all emotion classes with the progression of epochs. For the sad class, accuracy ranges from 95.633% to 95.813%, showing a gradual improvement over epochs. Similarly, the anger, tender, and happy classes also demonstrate incremental improvements in accuracy, with values ranging from 94.856% to 95.036%, 97.125% to 97.305%, and 96.989% to 97.169%, respectively. Precision, which measures the proportion of true positive predictions among all positive predictions, exhibits similar trends. There are slight increases in precision for all emotion classes as epochs increase. Specifically, precision values for the sad, anger, tender, and happy classes range from 94.856% to 95.036%, 94.353% to 94.533%, 94.867% to 95.047%, and 94.633% to 94.813%, respectively. Moving on to recall, which quantifies the proportion of true positive predictions among all actual positive instances, we observe varied trends across emotion classes. While the sad and anger classes demonstrate marginal increases in recall over epochs, the tender class exhibits fluctuating values, and the happy class shows consistent improvement. Recall values for the sad, anger, tender, and happy classes range from 93.969% to 94.149%, 94.526% to 94.706%, 93.895% to 94.075%, and 95.898% to 96.078%, respectively. The F1-score balances precision and recall, presents similar trends to precision and accuracy. There are gradual increases in F1-score across all emotion classes with the progression of epochs. Values for the sad, anger, tender, and happy classes range from 94.410% to 94.590%, 94.439% to 94.619%, 94.378% to 94.559%, and 95.261% to 95.441%, respectively.

Table 2. Comparative analysis of proposed and existing models for music emotion recognition

Music emotion recognition models	Values in %			
	Accuracy	Precision	Recall	F1-score
SVM + Relief	54.631	65.523	68.563	67.009
VGG16	56.457	63.147	72.963	67.701
Inception-v3	68.122	71.023	56.389	62.866
ResNet-18	74.357	69.856	63.255	66.392
DCNN [32]	82.541	79.856	61.872	69.723
CNN [33]	83.987	75.198	74.534	74.865
CNN+ Inception-GRU [31]	84.234	80.456	79.633	80.042
F-CNN+ELNN	96.241	94.767	96.612	95.681

Fig. 2 depicts the accuracy comparison of different models for music emotion recognition, providing insights into the performance across various approaches. Beginning with the SVM + Relief model, it achieves an accuracy of 54.631%, representing a baseline performance level. Moving forward, we observe a gradual increase in accuracy with subsequent models. VGG16 exhibits a modest improvement, with an increase of

approximately 1.826% compared to SVM + Relief. Inception-v3 demonstrates a more substantial increase, surpassing VGG16 by 11.665%. ResNet-18 continues this trend with an increase of approximately 6.235% over Inception-v3. The DCNN model, as reported in [32], shows a wise increase of 8.184% compared to ResNet-18, indicating significant advancements in accuracy. CNN further enhances performance with a wise increase of 1.446% over DCNN. Similarly, the CNN+ Inception-GRU model exhibits a wise increase of 0.247% over CNN. Remarkably, the F-CNN+ELNN model achieves the highest accuracy among all models, with an increase of approximately 11.007% over CNN+ Inception-GRU. An improvement underscores the effectiveness of the F-CNN+ELNN model in enhancing music emotion recognition accuracy.

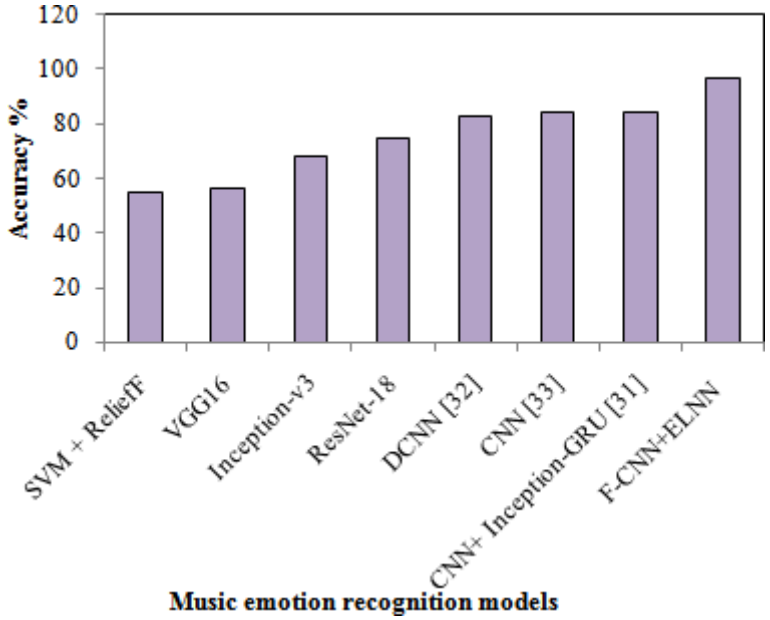


Fig. 2 Accuracy comparison of various models for music emotion recognition

Fig. 3 presents the precision comparison of different models for music emotion recognition, shedding light on the improvement in precision across various models. Initiating with the SVM + Relief model, it achieves a precision of 65.523%, serving as the baseline precision level. We observe fluctuations in precision across different models. VGG16 exhibits an improvement in precision compared to SVM + Relief, with decrease of approximately 2.376%. Inception-v3 shows notable increase in precision, surpassing VGG16 by 7.876%. ResNet-18 follows 1.167% compared to Inception-v3. The DCNN model, as reported in [32], shows an improvement in precision of 10.000% over ResNet-18, indicating advancements. CNN further enhances precision with an increase of 4.342% over DCNN. Similarly, the CNN+ Inception-GRU model exhibits an increase of 5.258% over CNN. Remarkably, the F-CNN+ELNN model achieves the highest precision among all models, with an increase of 14.311% over CNN+ Inception-GRU. This significant improvement highlights the efficacy of the proposed approach in enhancing precision for music emotion recognition.

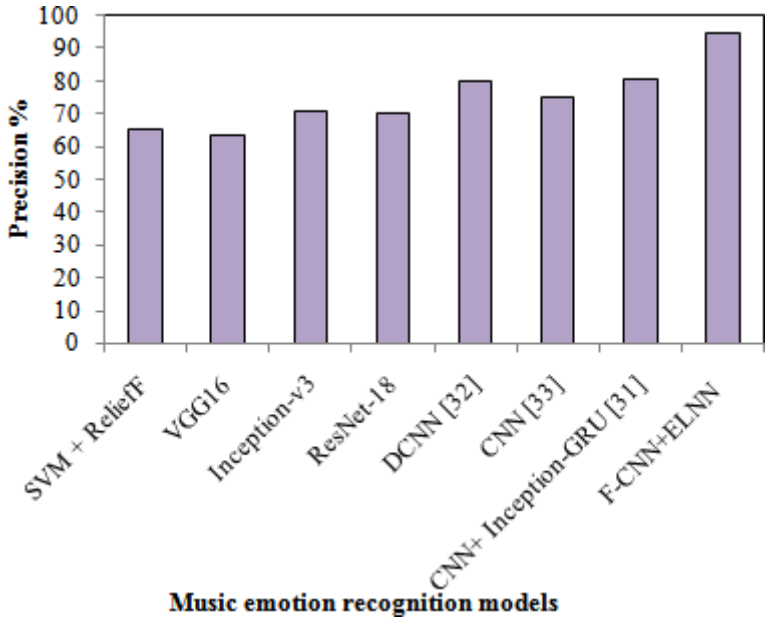


Fig. 3 Precision comparison of various models for music emotion recognition

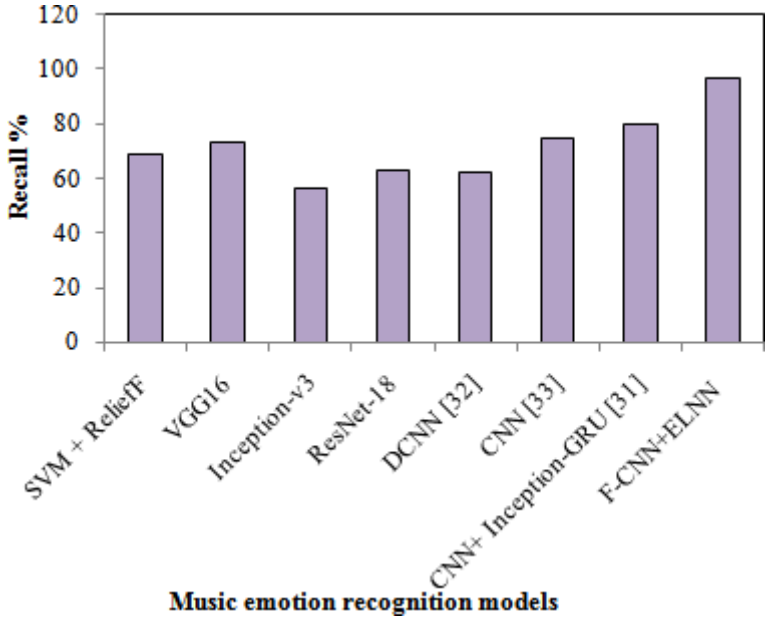


Fig. 4 Recall comparison of various models for music emotion recognition

Fig. 4 illustrates the recall comparison of different models for music emotion recognition, providing insights into the improvement in recall across various approaches. Commencing with the SVM+Relief model, it achieves a recall of 68.563%, serving as the baseline recall level. Subsequently, we observe fluctuations in recall across different models. VGG16 demonstrates a improvement in recall compared to SVM+Relief, with an increase of 4.4%. However, Inception-v3 exhibits a notable improvement in recall, falling short of VGG16 by 16.574%. ResNet-18 follows with an improvement of 6.866% over Inception-v3. The DCNN model, as reported in [32], shows an improvement in recall of 1.383% compared to ResNet-18. CNN further enhances recall with an improvement of 12.662% over DCNN. Similarly, the CNN+ Inception-GRU model exhibits an improvement of 5.099% over CNN. F-CNN+ELNN model achieves the highest recall among all models, with an improvement of 17.979% over CNN+ Inception-GRU. This substantial improvement highlights the effectiveness of the proposed approach in capturing emotional nuances in music with high recall compared to existing techniques. The significant increase in recall achieved by the F-CNN+ELNN model emphasizes its superiority in accurately identifying emotional states in music compared to existing techniques.

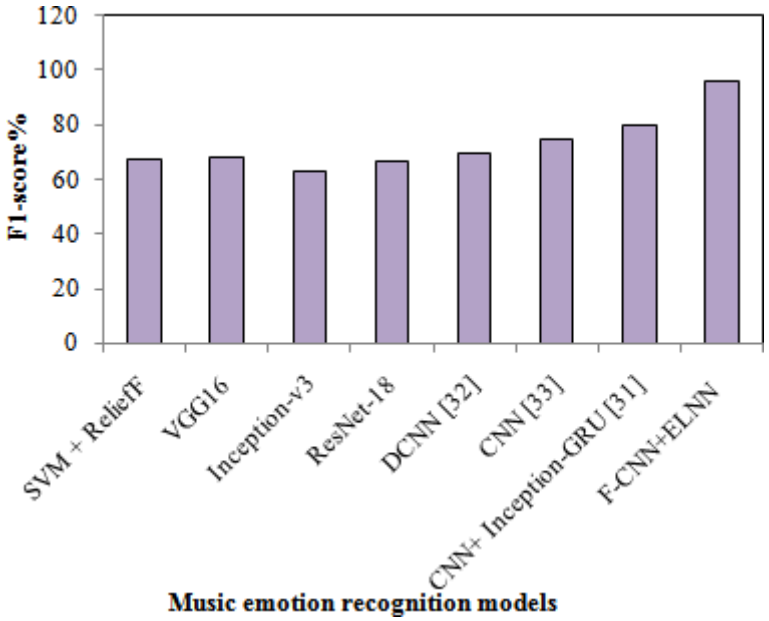


Fig. 5 F-measure comparison of various models for music emotion recognition

Figure 5 presents the F-measure comparison of various models for music emotion recognition, highlighting the improvement in F-measure across different approaches. Beginning with the SVM+Relief model, it achieves an F-measure of 67.9%, setting the baseline for comparison. VGG16 demonstrates a slight improvement in F-measure compared to SVM+Relief, with an increase of 0.692%. However, Inception-v3 exhibits a notable improvement in F-measure, falling short of VGG16 by 4.835%. ResNet-18 follows with an improvement of 3.526% over Inception-v3. DCNN model, as reported in [32], shows improvement in F-measure of 3.331% compared to ResNet-18. CNN further enhances F-measure with an improvement of 5.142% over DCNN. Similarly, the CNN+ Inception-GRU model exhibits an improvement of 5.177% over CNN. F-CNN+ELNN model achieves the highest F-measure among all models, with an improvement of 15.639% over CNN+ Inception-GRU. This substantial improvement highlights the effectiveness of the proposed model in achieving superior F-measure compared to existing techniques. The significant increase in F-measure achieved by F-CNN+ELNN model underscores its superiority in accurately recognizing music emotions compared to existing techniques.

5. Conclusion

The study introduces a hybrid soft computing approach for music emotion recognition, leveraging deep feature extraction and optimal feature selection techniques. Initially, a feeding convolutional neural network (F-CNN) layer is working to excerpt deep features from audio signals, complemented by the implementation of a modified wolf pack (MWP) cluster-based algorithm to optimize feature clustering and address challenges with the data dimensionality. An enhanced Laguerre neural network (ELNN) model is used for predicting emotional dimensions' levels. Through extensive experimentation on the Soundtrack dataset containing 360 sound samples, F-CNN+ELNN model shows exceptional performance. It achieved a maximum accuracy of 97.568%, representing an improvement of 12.476% over existing models. The model achieved a maximum precision of 94.767%, marking enhancements of 15.101% compared to existing models. The F-CNN+ELNN model achieved a maximum recall of 96.612%, indicating significant improvements of 17.574% over existing models. The maximum F1-score of the F-CNN+ELNN model reached 95.681%, marking significant enhancements of 16.344% compared to existing models. These results underscore the efficacy of F-CNN+ELNN model in advancing music emotion recognition, highlighting potential to outperform existing models across various performance metrics.

References

1. Orio, Nicola, and François Déchelle. "Score following using spectral analysis and hidden Markov models." 2001.

2. Chen, Yu-An, et al. "Linear regression-based adaptation of music emotion recognition models for personalization." 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.

3. Whissel, Cynthia M. "The dictionary of affect in language R. Plutchnik and H. Kellerman (Eds) Emotion: Theory, research and experience: vol 4, The measurement of emotions." (1989).

4. Martin, Rod A., et al. "Emotion perception threshold: Individual differences in emotional sensitivity." Journal ofResearch in Personality 30.2 (1996): 290-305.

5. Russell, James A. "A circumplex model of affect." Journal of personality and social psychology 39.6 (1980): 1161.

6. Gomez, Patrick, and Brigitta Danuser. "Relationships between musical structure and psychophysiological measures of emotion." Emotion 7.2 (2007): 377.

7. Yang, Li-Chia, Szu-Yu Chou, and Yi-Hsuan Yang. "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation." arXiv preprint arXiv:1703.10847 (2017).

8. Dong, Hao-Wen, et al. "Musegan: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment." arXiv preprint arXiv:1709.06298 (2017).

9. Laurier, Cyril, et al. "Music Mood Representations from Social Tags." ISMIR. 2009.

10. Wang, Ju-Chiang, et al. "Exploring the relationship between categorical and dimensional emotion semantics of music." Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. 2012.

11. T. Eerola, O. Lartillot, and P. Toiviainen, “Prediction of multidimensional emotional ratings in music from audio using multivariate regression models.” in ISMIR, 2009, pp. 621–626.

12. P. A. Wood and S. K. Semwal, “On exploring the connection between music classification and evoking emotion,” in Collaboration Technologies and Systems (CTS), 2015 International Conference on. IEEE, 2015, pp. 474–476.

13. A. Gabrielsson and E. Lindstrom, “The role of structure in the musical ¨ expression of emotions,” Handbook of music and emotion: Theory, research, applications, pp. 367–400, 2010.

14. W. H. O. G. Initiative and A. N. Overview, “WHO Global Initiative for childhood cancer – India responds,” *PediatricHematology Oncology Journal*, vol. 5, no. 4, pp. 145–150, 2020.

15. C. G. Lam, S. C. Howard, E. Bouffet, and K. Pritchard-Jones, “Science and health for all children with cancer,” *Science*, vol. 363, no. 6432, pp. 1182–1186, 2019.

16. Z. J. Ward, J. M. Yeh, N. Bhakta, A. L. Frazier, and R. Atun, “Estimating the total incidence of global childhood cancer: a simulation-based analysis,” *The Lancet Oncology*, vol. 20, no. 4, pp. 483–493, 2019.

17. K. Plummer, M. McCarthy, I. McKenzie, F. Newall, and E. Manias, “Experiences of pain in hospitalized children during hematopoietic stem cell transplantation therapy,” *Qualitative Health Research*, vol. 31, no. 12, pp. 2247–2259, 2021.

18. M. Van Schoors, J. De Mol, H. Morren, L. L. Verhofstadt, L. Goubert, and H. Van Parys, “Parents’ perspectives of changes within the family functioning after a pediatric cancer diagnosis: a multi family member interview analysis,” *Qualitative Health Research*, vol. 28, no. 8, pp. 1229–1241, 2018.

19. J. Livesley and T. Long, “Children’s experiences as hospital in-patients: voice, competence and work. Messages for nursing from a critical ethnographic study,” *International Journal of Nursing Studies*, vol. 50, no. 10, pp. 1292–1303, 2013.

20. H. Greenlee, M. J. DuPont-Reyes, L. G. Balneaves et al., “Clinical practice guidelines on the evidence-based use of integrative therapies during and after breast cancer treatment,” *CA: A Cancer Journal for Clinicians*, vol. 67, no. 3, pp. 194–232, 2017.

21. Huang, Chih-Fang, and Cheng-Yuan Huang. "Emotion-based AI music generation system with CVAE-GAN." *2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*. IEEE, 2020.

22. Wood, Pamela A., and Sudhanshu K. Semwal. "An algorithmic approach to music retrieval by emotion based on feature data." *2016 Future Technologies Conference (FTC)*. IEEE, 2016.

23. Huang, Chih-Fang, and Yin-Jyun Luo. "A study of the integrated automated emotion music with the motion gesture synthesis." *2011 IEEE International Conference on Computer Science and Automation Engineering*. Vol. 3. IEEE, 2011.

24. Weng, Jui-Feng, et al. "Constructing an Immersive Poetry Learning multimedia environment using ontology-based approach." *2008 First IEEE International Conference on Ubi-Media Computing*. IEEE, 2008.

25. Group, Indian Pediatric Oncology, CankidsKidscan Representing Civil Society, and Patient Groups. "WHO Global Initiative for childhood cancer–India responds." *PediatricHematology Oncology Journal* 5.4 (2020): 145-150.

26. Musliu, Arian, and Anne C. Frenzel. "Teachers' Emotions: Validation of the Teacher Emotions Scales in Albanian." *Education Research International* 2023 (2023).

27. Thamilarasi V., Roselin. R., "Application of Machine Learning in chest X-Ray images", International book & River publisher series in computing and information science and Technology, pp: 52-1,2023. ISBN : 9788770228114, 9788770228107.

28. Chen, Wei, and Guobin Wu. "A Multimodal Convolutional Neural Network Model for the Analysis of Music Genre on Children's Emotions Influence Intelligence." *Computational Intelligence and Neuroscience* 2022 (2022).

29. Hou, Yimin, and Shuaiqi Chen. "Distinguishing different emotions evoked by music via electroencephalographic signals." *Computational intelligence and neuroscience* 2019 (2019).

30. Qian, Qingfang, and Xiaofeng Chen. "A multi-modal convolutional neural network model for intelligent analysis of the influence of music genres on children's emotions." *Computational Intelligence and Neuroscience* 2022 (2022).

31. Ma, Lina. "Algorithmic application of evidence theory in recommender systems." *Scientific Programming* 2022 (2022).

32. Han, X., Chen, F. and Ban, J., 2023. Music Emotion Recognition Based on a Neural Network with an Inception-GRU Residual Structure. Electronics, 12(4), p.978.

33. Hassan Yacin. (2026). AI-Driven Energy-Aware Routing Protocol for Scalable Wireless Sensor Networks in 5G-Enabled IoT Environments. *Journal of Wireless Sensor Networks and IoT*, *3*(2), 11-18.

34. Sarkar, R., Choudhury, S., Dutta, S., Roy, A. and Saha, S.K., 2020. Recognition of emotion in music based on deep convolutional neural network. Multimedia Tools and Applications, 79, pp.765-783.

35. Chaudhary, D., Singh, N.P. and Singh, S., 2021. Development of music emotion classification system using convolution neural network. International Journal of Speech Technology, 24, pp.571-580.