Journal of Scientific & Industrial Research
Vol. 84, March 2025, pp. 308-323
DOI: 10.56042/jsir.v84i03.13751

DUALBIGRU-UCSA: Deep Learning based Music Emotion Recognition Model

Szeto Chung Man[1], Alok Kumar[2]*, Ajay Tiwari[3], Prateek Srivastava[4], Deepak Kumar Verma[2], Pushpa Mamoria[5], Vineeta Singh[2]*, Chandra Shekhar Kumar[6], Amit Seth[7], Kapil Joshi[8] & Vandana Dixit Kaushik[9]

[1]School of Music and Recording Arts, Communication University of China, Beijing, China

[2]Department of Computer Science and Engineering, [3]Department of Electronics & Communication Engineering,

[4]Department of Information Technology, [5]Department of Computer Applications, School of Engineering and Technology (UIET), [6]Department of Physiotherapy, School of Health Sciences, Chhatrapati Shahu Ji Maharaj University, Kalyanpur, Kanpur, Uttar Pradesh 208 024, India

[7]Department of Information Technology, Galgotias College of Engineering and Technology, Knowledge Park II, Greater Noida, Uttar Pradesh 201 310, India

[8]Department of Computer Science & Engineering, Uttaranchal Institute of Technology (UIT), Uttaranchal University, Dehradun 248 007, Uttarakhand, India

[9]Department of Computer Science and Engineering, Harcourt Butler Technical University Kanpur, Nawabganj Uttar Pradesh 208 002, India

Received 18 September 2024; revised 25 December 2024; accepted 28 December 2024

Music Emotion Recognition (MER) is a process to classify emotions perceived in a given piece of music with computational models. There are several problems regarding existing models, due to subjective perception of emotions and individual differences and culture diversity. To overcome these challenges, we developed a Dual Bidirectional Gated Recurrent Unit with Unified Contextual Shuffle Attention Fusion (DualBiGRU-UCSA) model. Here, the primary contribution lies in the practical implementation of bidirectional and gated recurrent units along with developed attention mechanisms to address the requirements for understanding and perceiving complex musical features. Using Bidirectional GRUs, the model taps the information of past and future contexts of music sequences in addition to refining the features of temporal dynamics and feelings. The final model's performance enhancements involve the integration of bidirectional GRU outputs to the UCSA module through paying much attention and shifting through the feature representations, the module consisting of Shuffle Attention and Multi-Head Location-Aware Attention performs by reducing the unimportant feature representations while enhancing the important patterns and contextual cues. The proposed model performs better in terms of high accuracy, f1-score, negative predictive values, positive predictive values and recall of 96.28%, 96.32%, 96.26%, 96.60% and 96.27% respectively as compared to recent State-of-the-Art techniques.

Keywords: Dual bidirectional gated recurrent unit, Mel frequency cepstral coefficients, Music emotion recognition, Unified contextual shuffle attention fusion, Weighted categorical cross-entropy,

Introduction and Literature Review

Human emotions are a major component of music, and they play a crucial role in both understanding and appreciation of music and transmission of emotions.[1] The quantity of digital music is expanding quickly due to AI advancements and Internet technologies.[2] MER has garnered significant interest from specialists and scholars in the fields of music database administration, music retrieval, recommendation, and music therapy.[3] In the fields of education, psychologically oriented counseling, music therapy, and other fields, music serves a variety of purposes.[4] The subjective nature of music emotion understanding is strongly related to personal traits like age, profession, family environment, and background.[5] Recent advancements in neuroscience enable researchers to compute the impact of music on the human brain in addition to having a deep understanding of the concepts.[6]

Emotional disorder treatment and music therapy are two areas where active applications of music emotion detection systems are found.[7] An interdisciplinary project involving signal processing, machine learning, auditory perception principles, psychology, cognitive science, and musicology is determining the emotional contents of music through computations.[8] Emotion identification from music signals is a challenging task. People may experience diverse feelings when listening to the same song, contributing to the difficulty in

*Author for Correspondence
E-mail: vineeta.singh.cs@gmail.com; alok@csjmu.ac.in

interpreting emotion perception.[9] The performance evaluation of a music emotion identification system becomes very challenging due to this subjectivity issue. While different people may use different characteristics to characterize the same feeling, leading to challenges to define emotions in a universal way. Ultimately, the exact mechanism by which music evokes human emotions remains a mystery, and the inner workings of music's emotional impact on the listener remain incompletely understood.[10] The psychological emotional technique is frequently employed in numerous studies in this discipline. There are two primary categories into which the automated methods in use today for identifying emotions in music can be divided: categorically and dimensionally.[11] Discrete models, classify emotions using single words or phrases (e.g. happy, sad, angry, calm). Different emotional states are positioned in a space that is converted into two-dimensional emotional states in dimensional models.[12]

Within the Thayer's two-dimensional emotion model, valence and stimulation levels are quantified along horizontal and vertical axes, respectively.[13] Using these widely used models, the music is emotionally classified into several classifications.[14] Rhythm, tone, and harmony and other characteristics indicating the acoustic content of music are retrieved in the conventional music emotion detection methods.[15] The challenges associated with extracting features from music recordings and the lack of assurance regarding the relationship between recovered features and musical emotions are the drawbacks of the conventional methods.[16] Therefore, a new model is required to address the current issues. The 1-D CNN-based Inception CRU residual model's transfer learning capability was not fully utilized, resulting to increase in sample data for a shorter training time.[17] The CNN-based auto encoder + BiLSTM model lacked robustness, and the impact of masking span on performance was not examined.[8] The BLNN model showed poor recognition efficiency and accuracy, and it was unable to assess multi-modal information.[6] Verse1, Chorus, Verse2 Structure demonstrated poor accuracy and interpretability in MER systems and lacked attention mechanisms.[18] The accuracy was impacted by the FFA-BiGRU model poor classification and limitations in a variety of datasets due to varied emotional reactions to the same music.[19] The muSi-ABC architecture was first presented by Yang.[4] Nevertheless, the generality was diminished since more modal information was overlooked. An embedding-based method for MER was presented by Takashima *et al.*[11] still unable to obtain up on the preexisting feature representations. Chen[12] introduced a Long Short-Term Memory (LSTM) network model for multimodal music emotion analysis.

Deam dataset[20] includes 58 songs from the 2015 evaluation set, 1000 songs from the 2014 evaluation set, and 744 songs from the 2014 development set. Wang *et al.*[21] presented Dual BiGRU-CNN-based sentiment classification method combining global and local attention. Recently different emotion recognition models were proposed.[22–26] Current approaches used for MER pose a number of challenges such as insufficient ability to model temporal patterns and inability to model the variations in emotions in sequences of music. Traditional models may also fail at fitting and encoding temporal features in both ways, thus having limited knowledge of emotions, and providing relatively low accuracy. These limitations are overcome through the proposed DualBiGRU-UCSA model involving dual bidirectional GRUs that improve the temporal sequence analysis and a UCSA module optimizing the feature understanding.

The input for identifying musical emotions is based on data gathered from[20] and is expressed as:

$$F_d = \sum_{n=1}^{t} F_n \quad \ldots (1)$$

In this case, F_n represents the number of data in the dataset, which spans from 1 to t, and F_d represents the database.

Preprocessing of the input audio signal is an important step in improving the quality of the raw input to eliminate noise while preserving the useful frequency band. In this research preprocessing is done with the use of a bandpass filter

$$F_d^* = F_d * h(t) \quad \ldots (2)$$

Here, the input signal is denoted as F_d and the band pass filter impulse response is denoted as $h(t)$, further feature extraction is one of the important steps in this research. The final output is afterwards employed to compute the loss utilizing a hybrid loss function. The hybrid loss consists of Weighted Categorical Cross-Entropy and other auxiliary loss terms, proving that it correctly learns temporal and spatial fields. The loss is expressed as

$$L = \alpha.L_{WCCE} + \beta.L_{auxilary} \quad \ldots (3)$$

Here, weighted Categorical Cross-Entropy loss is abbreviated as L_{WCCE} and the auxiliary loss are

abbreviated as $L_{auxilary}$ while the weighting factors are defined by α and β. .

System Model

The input music data is processed to go through a preprocessing phase whereby noise is filtered out and the signal is normalized for proper comparison. Afterwards the system moves to feature extraction phase where attributes such as rhythm, melody, tempo and spectral characteristics of music data are extracted. These extracted features are then passed to a DL model for emotion recognition. In this phase the model is learnt on a labeled dataset to be able to map particular quantitative feature values to some emotional states. Further the model is trained using the signals, it is tested using another different test signal. Finally the trained model processes the test signal and categorizes it into any of the defined emotions like joy, anger or sorrow. The output of the system is the anticipated emotion of the input music which serves as the understanding of the emotional intent of the musical piece. System model for the Music emotion recognition system is shown in Fig. 1.

Proposed Methodology for Music Emotion Recognition Using Dual Bidirectional Gated Recurrent Unit with Unified Contextual Shuffle Attention Fusion Model

The goal of this research is to identify different emotions that are present in the music signals, such as joy, sadness and anger. First, the preprocessor receives the input audio signal that is gathered from the DEAM dataset.[20] Here, the undesired noise is eliminated to improve the raw input's quality. The stage of feature extraction takes the enhanced signal and extracts features such as spectral contrast; spectral centroid, zero-crossing rate, bandwidth, entropy, VGG16 features, Mel Frequency Cepstral Coefficients (MFCC), and Short-Time Fourier Transform (STFT) that extract the frequency content changes of non-stationary signals. The Bidirectional GRU, which consists of two GRUs with input and forget gates, one of which take input in the forward direction and the other in the backward direction is then fed the extracted signal features. By allowing the signal information to be remembered or forgotten over time, this aids in modeling the sequential data. The model's losses are quantified by the hybrid loss of weighted categorical cross-entropy. The hybrid multi-head location-aware attention and shuffle attention reduce workload of the long sequences and increase model performance. To test a trained model and assess its performance in detecting emotions, a range of signals are used. The proposed Music emotion recognition model is illustrated in Fig. 2.

Input from DEAM Dataset

The input for identifying musical emotions is based on data gathered from DEAM dataset[20] and is expressed mathematically as follows.

$$F_d = \sum_{n=1}^{t} F_n \quad \ldots (4)$$

Preprocessing using Bandpass Filter

Preprocessing of the input audio signal is an important step in improving the quality of the raw input to eliminate noise while preserving the useful frequency band. Here preprocessing is done with the use of a bandpass filter allowing signals falling within

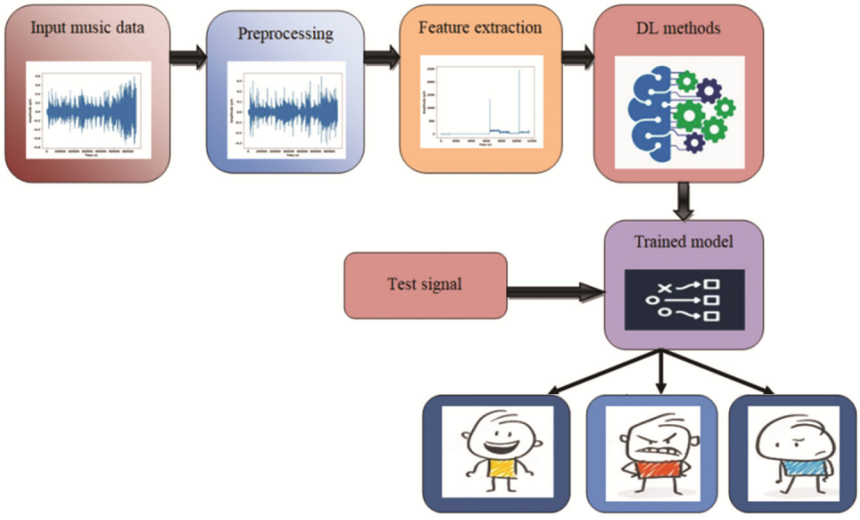

Fig. 1 — System model for the music emotion recognition

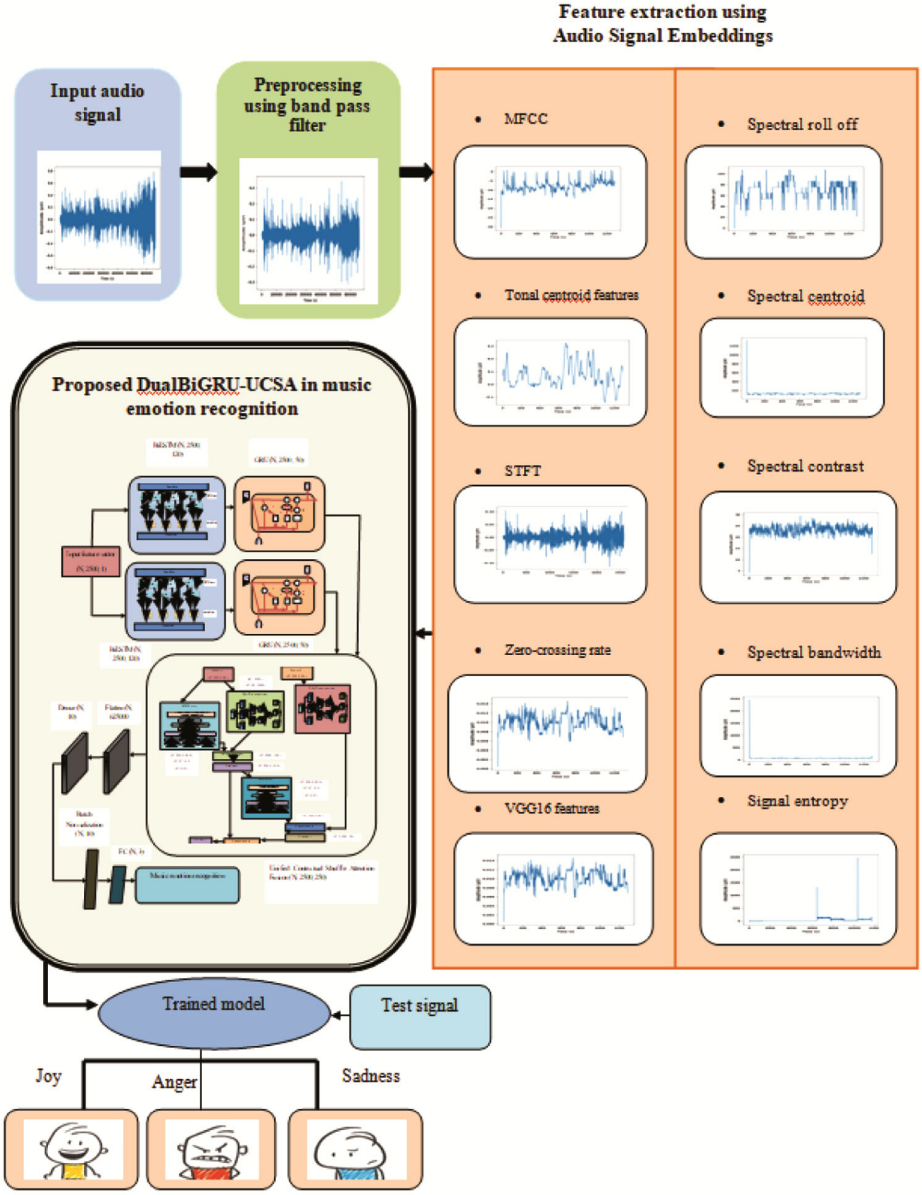

Fig. 2 — Architecture for the proposed music emotion recognition

a given range of frequencies while rejecting the other frequencies. The bandpass filter helps to filter out most of the unwanted noise and concentrate on the range of frequencies which can be most useful in relation to emotion detection in music signals. The sample rate obtained from the preprocessed signal is 22050. The preprocessed output is provided in Eq. 2.

Feature Extraction Using Audio Signal Embeddings

It is one of the important steps here, through which the necessary characteristics of the preprocessed signals are obtained to evaluate and classify the emotions incorporated into the music. The preprocessed audio signal represented by F_d^* is then subjected to several transformations with the aim of extracting features that incorporates both the temporal and spectral domain characteristics of the signal.

The process of obtaining the harmonic content of music signals which is essential for identifying emotions in the audio is facilitated by Tonal Centroid characteristics. A 6-dimensional feature vector is used to represent the tonal centroid, which is defined as a projection of the chords along the minor, major, and fifth circles. The output dimension obtained from the tonnetz feat is 1292.

The deep network features of the VGG16 architecture are utilized to improve the accuracy of emotion identification, making it a potent tool for feature extraction from audio stream. With 13 convolution and 3 full-connected layers, the VGG16 pre-trained deep network architecture is a deep network, included in its total number of layers are the Maxpool, Full connected, Relu, Dropout, and SoftMax layers, thus forming the categorization layer is the final step. Maxpooling is used to reduce dimension after a few of the convolution layers. Compared to Alex Net, which uses larger filters, smaller 3 × 3 filters are utilized. The output dimension obtained from the vGG 16 feat is 100.

MFCC have been applied successfully in automatic speech recognition which has gained a lot of traction in the MIR society, where they are successfully applied to the classification of musical genres and perceptually significant groupings, like moods and perceived complexity. Consider a signal's Fourier transform applied on a windowed snippet.

$$S(f)=F\left(F_d^*(t).w(t)\right) \quad \ldots (5)$$

In this case, the windowing function is represented by $w(t)$, the frequency domain representation is designated by $S(f)$, and the Fourier transform is represented by F. Using triangle overlapping windows, map the powers of the spectrum that were acquired above onto the Mel scale.

$$pitch(Mel)=1127.0148\left(1+\frac{f}{700}\right) \quad \ldots (6)$$

Record the power logs for each of the Mel frequencies.

$$\log F_j=\log\left(F_j\right)$$

Apply the discrete cosine transform to the list of Mel log powers, considering it as a signal.

$$d_j=\sum_{k=1}^{N_f}\log(F_k)\cos\left[j\left(k-\frac{1}{2}\right)\frac{\pi}{N_f}\right], 1\le j\le N_d \quad \ldots (7)$$

Spectral energy recorded in the critical band of the k^{th} Mel filter is denoted by F_k in this case, and the total Mel filters is represented by N_f (typically N_f = 24). N_d is the number of Cepstral coefficients d_j that are acquired from each frame; in this case, N_d = 13. The MFCCs' derivation procedure is shown in Fig. 3. The output dimension obtained from the MFCC feet is 1292.

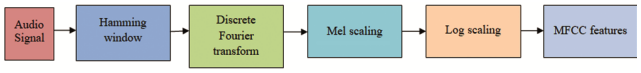

Fig. 3 — MFCC derivation block diagram

The number of times a signal changes signs is known as the zero crossing rate. Voiced segments typically have lower zero crossing rates than unvoiced portions. To extract the zero crossing rate apply Eq. (8).

$$z_j=\frac{1}{P}\sum_{p=1}^{P-1}\left|sign\left(F_d^*(p)\right)-sign\left(F_d^*(p-1)\right)\right| \quad \ldots (8)$$

Here, zero crossing rate is denoted as z_j, total number of samples present in the frame is denoted as P, $sign\left(F_d^*(p)\right)$ is the sign function applied to the signal $F_d^*(p)$. The output dimension obtained from the zero crossing rates is denoted as 1292.

$$sign\left(F_d^*(p)\right)=\begin{cases}1, & if\ F_d^*(p)>0\\ 0, & if\ F_d^*(p)=0\\ -1, & if\ F_d^*(p)<0\end{cases} \quad \ldots (9)$$

Here, $sign\left(F_d^*(p)\right)-sign\left(F_d^*(p-1)\right)$ counts a zero crossing when there is a change in signal sign between successive samples.

First apply the Fourier transform for preprocessed signal F_d^* to get its frequency spectrum $R(L)$. Here t denotes the time.

$$R(L)=F\left\{F_d^*(t)\right\} \quad \ldots (10)$$

Here, frequency bin index is denoted as L, signal magnitude is denoted as $R(L)$

Spectral Rolloff (SR) analysis identifies the frequency at which 85% of the signal's magnitude is concentrated, highlighting the signal's spectral skewness. This metric is essential for distinguishing different types of audio signals. An output dimension of 1292 from Spectral Roll off indicates a detailed analysis, capturing extensive spectral features across multiple frames or windows, enhancing the understanding of the signal's frequency distribution.

$$\sum_{l=0}^{L_{roll}}|R(L)|=0.85\sum_{l=0}^{\frac{N}{2}}|R(L)| \quad \ldots (11)$$

Here, the frequency bin is denoted as L_{roll}, total frequency bins are denoted as N, $R(L)$ denotes the spectrum magnitude.

First apply the Fourier transform for preprocessed signal F_d^* to get its frequency spectrum $SF_d^*(f)$. Here t denotes the time.

$$SF_d^*(f) = F\left\{F_d^*(t)\right\} \quad \ldots (12)$$

Here, the magnitude of the signal at each frequency is denoted as $SF_d^*(f)$

The baricenter of the spectrum, known as the spectral centroid, is determined by summing the weights of all the frequencies in the signal and using their magnitudes as weights. The output dimension obtained from the spectral centroid feat is denoted as1292.

$$SC = \frac{\sum_{n=0}^{N-1} f(n)\left|SF_d^*(n)\right|}{\sum_{n=0}^{N-1}\left|SF_d^*(n)\right|} \quad \ldots (13)$$

Here, the central frequency is denoted as $f(n)$, total frequency bins are denoted as N, spectrum magnitude derived from F_d^* is denoted as $SF_d^*(n)$.

$$SC = \frac{\sum_{n=0}^{N-1} f(n)\left|F\left\{F_d^*(t)(n)\right\}\right|}{\sum_{n=0}^{N-1}\left|F\left\{F_d^*(t)\right\}(n)\right|} \quad \ldots (14)$$

Here, the Fourier transform of F_d^* is denoted as $F\left\{F_d^*(t)(n)\right\}$,

An audio signal's spectrum properties, more precisely its relative spectral distribution, can be represented using a feature called octave-based spectral contrast. In experiments with music type classification difficulties, the feature performs better than one of the often used features, MFCC, in distinguishing between various music types. The output dimension obtained from the spectral contrast feat is 1292.

The second central moment of variance is known as the spectrum spread/instantaneous spectral bandwidth. Squared deviation of the distribution from its mean value is the variance, a measure of the dispersion or spread of the spectrum that is always positive. The equation used to calculate is given below. The output dimension obtained from the spectral band width is 1292.

$$SB = \sum_{n=0}^{N-1} SF_d^*(n)(f(n) - SC)^2 \quad \ldots (15)$$

Here, the central frequency at n^{th} bin is denoted as $f(n)$ and spectral centroid is denoted as SC.

First apply the Fourier transform for preprocessed signal F_d^* to get its frequency spectrum $SF_d^*(f)$. Here t denotes the time.

$$SF_d^*(f) = F\left\{F_d^*(t)\right\} \quad \ldots (16)$$

The signal spectrum's Shannon's entropy is used to calculate spectral entropy. To get a value that is independent of the windowing length, divide the entropy by the logarithm of the spectrum's length.

$$SE = -\frac{\sum_{l=1}^{L} SF_d^*(l).\log\left(SF_d^*(l)\right)}{\log(N)} \quad \ldots (17)$$

Here, the spectrum magnitude at the l^{th} frequency bin is denoted as $SF_d^*(l)$, total number of frequency bins considered is denoted as l

$$SE = -\frac{\sum_{l=1}^{L} F\left\{F_d^*(t)\right\}(l).\log\left(F\left\{F_d^*(t)\right\}(l)\right)}{\log(N)} \quad \ldots (18)$$

Here, the Fourier transform of $F_d^*(t)$ is denoted as $F\left\{F_d^*(t)\right\}$ which provides the spectral magnitude $SF_d^*(l)$.

Short term (or time) Fourier transformations (STFT) are the preferred method for processing these signals, since the signal is windowed into brief intervals using STFT, which may be presumed that these signal segments are stationary. The breadth of this window must coincide with the signal segment where the dc component is assumed to remain constant and the signal is stationary. The window function is present at the start of the signal, and the signal's product with the window function is computed. After that, this product is regarded as merely another signal, for which the computation of Fourier transforms (FT) is required. This window would be moved to a new location, the signal would be multiplied, and the product's FT would be obtained. The output dimension obtained from the STFT feat is 2584.

$$STFT\left\{F_d^*\right\}(t,\omega) = \int_{-\infty}^{\infty} F_d^*(\tau)\omega(t-\tau)e^{-j\omega\tau}\,d\tau \quad \ldots (19)$$

Here, the window function is denoted as $\omega(t)$, angular frequency is denoted as ω and the time variable is denoted as τ.

Feature Concatenation

Finally Feature concatenation involve concatenating all the extracted features such as tonal centroid features, VGG16 features, MFCC, ZCR, SR, spectral centroid, spectral contrast, spectral bandwidth, spectral entropy and STFT which extracts the frequency content changes of non-stationary

signals and forms a final feature vector C. The final output dimension obtained from the feature concatenation is (N, 2500, 1).

$$C = \begin{bmatrix} Tonal // VGG16 // MFCC // zerocrossing // spectral\,roll\,off // spectral\,contrast // \\ spectral\,bandwidth // spectral\,entropy // STFT \end{bmatrix} \quad \ldots (20)$$

Dual Bidirectional GRU with Unified Contextual Shuffle Attention Fusion in Music Emotion Recognition

The final feature concatenated output (N, 2500, 1) forms the input for the music emotion recognition. DualBiGRU-UCSA is aimed at boosting the MER system by adopting the new deep learning methodologies. First, there is sequence data that is processed using BiLSTM layers to capture past and next context information of the sequences of music. The outputs of each of the BiLSTM layers are of 120 features and they are combined together in order to come up with 240 features that would represent the input sequences extensively and comprehensively. These representations are then passed through two Gated Recurrent Units (GRUs) each of which outputs 50 features then concatenated to enhance the temporal properties of this data. Subsequently, UCSA module is utilized which divides the outputs of the GRU into segments and performs channel and spatial attention to pay more attention to the relevant features and add the contextual information. The Shuffle Attention mechanism arranges the features for the capture of multi-aspect context information while the MHLAA improves this having more concentration on high significant patterns and relationship found in data. The outputs from above attention mechanisms are concatenated together and then fed to dropout layers for model regularizations. Finally, integrating BiLSTM, GRU, attention-based mechanisms, the proposed DualBiGRU-UCSA model improves the identification of the emotional content of music.

As for the input data, there is an input sequence with length 2500 and 1 feature dimension, which enters the proposed model and is passed between two BiLSTM layers that conduct forward and backward temporal analysis. The two BiLSTM layers return feature maps of size (N, 2500, 120) the features extracted from both directions. These outputs are concatenated together to give overall feature vector C with shape of (N, 2500, 240) thus containing a good temporal information. This concatenated vector is passed through two GRU layers that have a dimensionality of 50 where the outputs are in the form of dimensionality of (N, 2500, 50). The outputs obtained from the GRU layers are again concatenated in order to have the required feature representation UCSA of dimensions (N, 2500, 100). This final concatenated feature vector is then passed on to the UCSA module where the spatial and channel wise dependencies are again fine tuned to be more precise in terms of temporal features and attention mechanisms are used to not only retain but also to reinforce the temporal nature of the input data.

$$\overrightarrow{h_i B1} = ForwardB1\left(C, \overrightarrow{h_{t-1}}\right) \quad \ldots (21)$$

$$\overleftarrow{h_i B1} = backwardB1\left(C, \overleftarrow{h_{t-1}}\right) \quad \ldots (22)$$

$$h_i B1 = \overrightarrow{h_i B1} + \overleftarrow{h_i B1} \quad \ldots (23)$$

Here, the BiLSTM 1 is denoted as $B1$. The BiLSTM 1 concatenated output forms the output dimension of $(N,2500,120)$

$$\overrightarrow{h_i B2} = ForwardB2\left(C, \overrightarrow{h_{t-1}}\right) \quad \ldots (24)$$

$$\overleftarrow{h_i B2} = backwardB2\left(C, \overleftarrow{h_{t-1}}\right) \quad \ldots (25)$$

$$h_i B2 = \overrightarrow{h_i B2} + \overleftarrow{h_i B2} \quad \ldots (26)$$

The BiLSTM 2 concatenated output forms the output dimension of $(N,2500,120)$. The outputs from the two BiLSTM layers are concatenated to form the feature vector F :

$$F1 = h_i^{B1} \quad \ldots (27)$$

$$F2 = h_i^{B2} \quad \ldots (28)$$

$$F = F1 \oplus F2 \quad \ldots (29)$$

The concatenated feature vector F is then fed into two GRU layers for further temporal refinement. The outputs from the two GRU layers are concatenated to form the final input for the UCSA module:

$$T = h^{GRU1} \oplus h^{GRU2} \quad \ldots (30)$$

Shuffle Attention Module is only good in encoding the spatial relationship while it fails when it comes to capturing channel-wise correlations which might result in loss of some important contextual information across the channels. At the same time, MHLAA can provide outstanding results in channel-wise dependence extraction utilizing attention

mechanisms while sometimes can ignore the spatial structure of features that is often crucial for data understanding. Moreover, either of the models alone may cause additional computational expense and operational complications, while sharing few characteristics of the other. To overcome these shortcomings, UCSA model is constructed after integrating the advantages of both Shuffle Attention and MHLAA. Such integration is beneficial to UCSA as it helps in capturing both spatial and channel-wise dependencies making it more informative. Thus, UCSA system integrates these two mechanisms and eliminates the drawbacks carried by individual approach improving the result of neural network for the further effective processing of the complicated information. In the proposed DualBiGRU-UCSA model, the UCSA Module takes in advanced attention mechanisms to enhance the feature representation learned from the GRU layers. First, the module takes two GRU layers' outputs, and each has the size of (N, 2500, 50) where N is the batch size, 2500 is the number of sequences, and 50 is the feature space. These outputs are inputted directly into the Shuffle Attention mechanism to divide the input into H groups, where each group Y_i has dimensions $\Re^{\left(\frac{C}{H}\right)\times W\times I}$. The output of all these attentions is added together to form the shuffle attention output stacked with the original shape of (N, 2500, 50). Similarly, the MHLAA works on this feature set, calculates the attention scores using the scaled dot-product operation, apply SoftMax function on the calculated attention scores and finally obtain the attention output. This attention output is of size (N, 2500, 50) where it combines multiple heads of attention in order to capture different contextual relations between the elements in the sequence data. The outputs of the MHLAA and Shuffle Attention are then used in later steps of the proposed model to perform more processing and feature representation refinement.

$$Y=[Y_1,Y_2,.....Y_H] \text{ with each } Y_i\in\Re^{\left(\frac{C}{H}\right)\times W\times I} \quad \ldots (31)$$

$$Y_{i1}^0=\sigma\left(W_1.avgpool\left(Y_{i1}\right)+B_1\right)\Theta Y_{i1} \quad \ldots (32)$$

$$Y_{i2}^0=\sigma\left(W_1.GN\left(Y_{i2}\right)+B_2\right)\Theta Y_{i2} \quad \ldots (33)$$

$$SAO=applyshuffleattention\left(Y_{suffle}\right) \quad \ldots (34)$$

$$Attention_{score}=soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad \ldots (35)$$

$$AO=Attentionscore.V \quad \ldots (36)$$

$$ConOut=con\left(MHLAA,shuffleattention\right) \quad \ldots (37)$$

$$DO1=dropout\left(ConOut\right) \quad \ldots (38)$$

$$MHLAAout=\text{Apply MHLAA}\left(DO1\right) \quad \ldots (39)$$

Here, the input feature map Y split into H group. C denotes the number of channels, feature map width is denoted as W, height is denoted as I, Y_i denotes the sub feature map with reduced channels. Learnable weight metrics are denoted as W_1, W_2, activation functions are denoted as σ, bias terms are denoted as B_1 and B_2, group normalization applied in the feature map is denoted as Y_i. N denotes the number of samples. Q denotes the query matrix, K denotes the key matrix, key vector dimension is denoted as d_k, transpose of key matrix is denoted as , attention output is denoted as AO, value matrix is denoted as V.

In DualBiGRU-UCSA model, the most significant processing stages are Concatenation and Dropout which helps in integrating and normalizing the features learnt. Firstly, the results obtained from the MHLAA and Shuffle Attention mechanisms are combined into the concat output that have dimensions of (N, 2500, 100). This concatenated output goes through a dropout operation to overcome the problem of over fitting and it produces a Dropout Output 1 with the same dimension. After this, the MHLAA is implemented to Dropout Output 1 to derive the MHLAA Final Output of size (N, 2500, 120). At the same time, the output from the second GRU layer is passed through the Shuffle Attention mechanism as a result of which, we obtain Second GRU Shuffle Attention Output of dimensions (N, 2500, 50). This output is then concatenated with the MHLAA Final Output as the Final Concatenated Output in the dimensions of (N, 2500, 150) which is again of the same dimensions known as Dropout Output 2 by applying a dropout to this final concatenated output. Finally, Dropout Output 1 and Dropout Output 2 are concatenated and form the Final Output with dimension of (N, 2500, 250). This final output incorporates the elements of variances in attention and includes Dropout layers making the model less prone to over fitting for further processing by other layers.

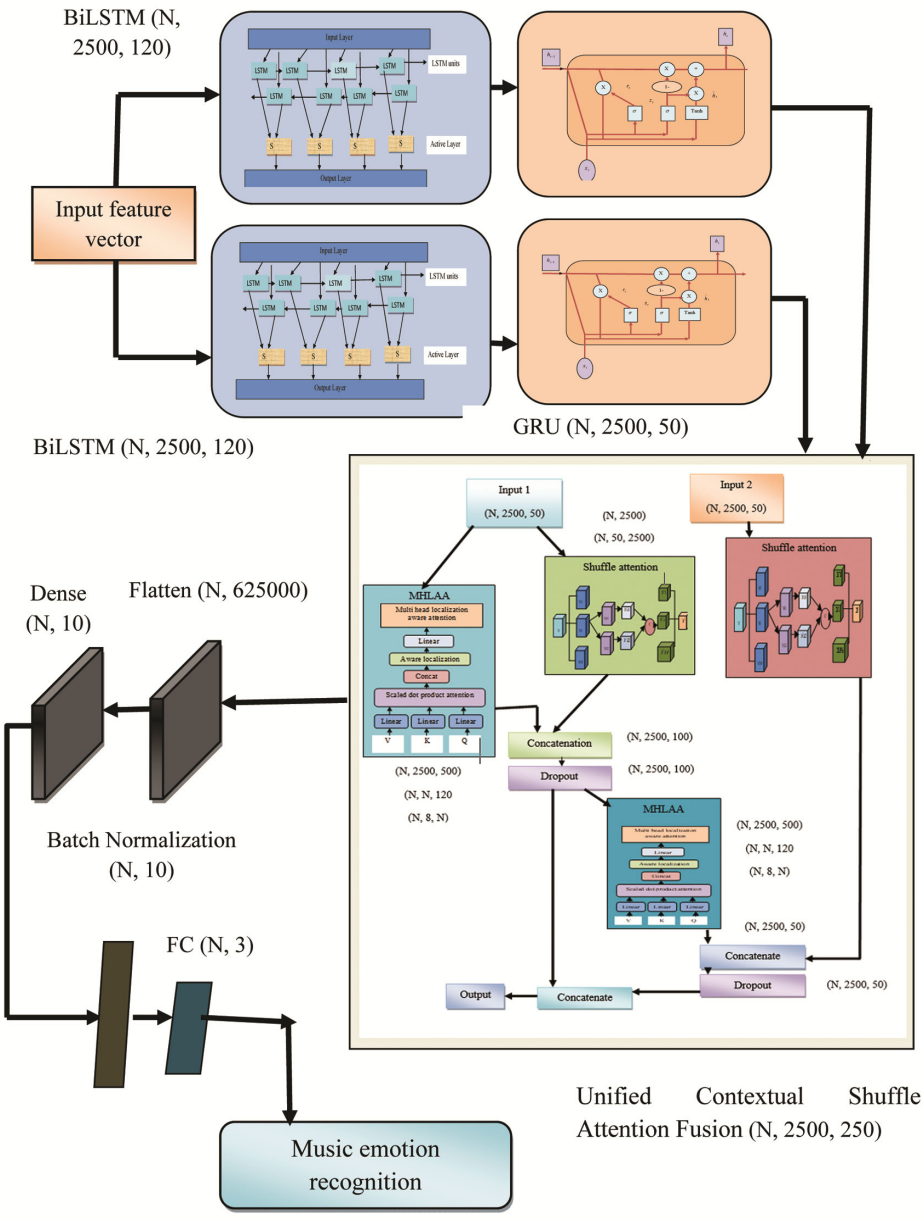

Fig. 4 — Proposed DualBiGRU-UCSA model architecture

Then passed through the shuffle attention and has an output of $^{Y_{GRU2}}$.

$$FO = concat(\text{second GRU shuffle attention output}) \quad \ldots (40)$$

Final concatenated output= $\in \Re^{(N,2500,150)}$

$$DO2 = dropout(FO) \quad \ldots (41)$$

$$FO = Concat(DO1, DO2) \quad \ldots (42)$$

The final output is afterwards employed to compute the loss utilizing a hybrid loss function. The hybrid loss consists of Weighted Categorical Cross-Entropy and other auxiliary loss terms, proving that it correctly learns temporal and spatial fields. The loss is expressed as in Eq. 3. The architectures for the proposed DualBiGRU-UCSA model and Unified Contextual Shuffle Attention Fusion are illustrated at Fig. 4 and 5.

Results & Discussion

To carry out the musical emotion recognition experiment, a python program is being executed on a Windows 10 OS with 8 GB of memory.

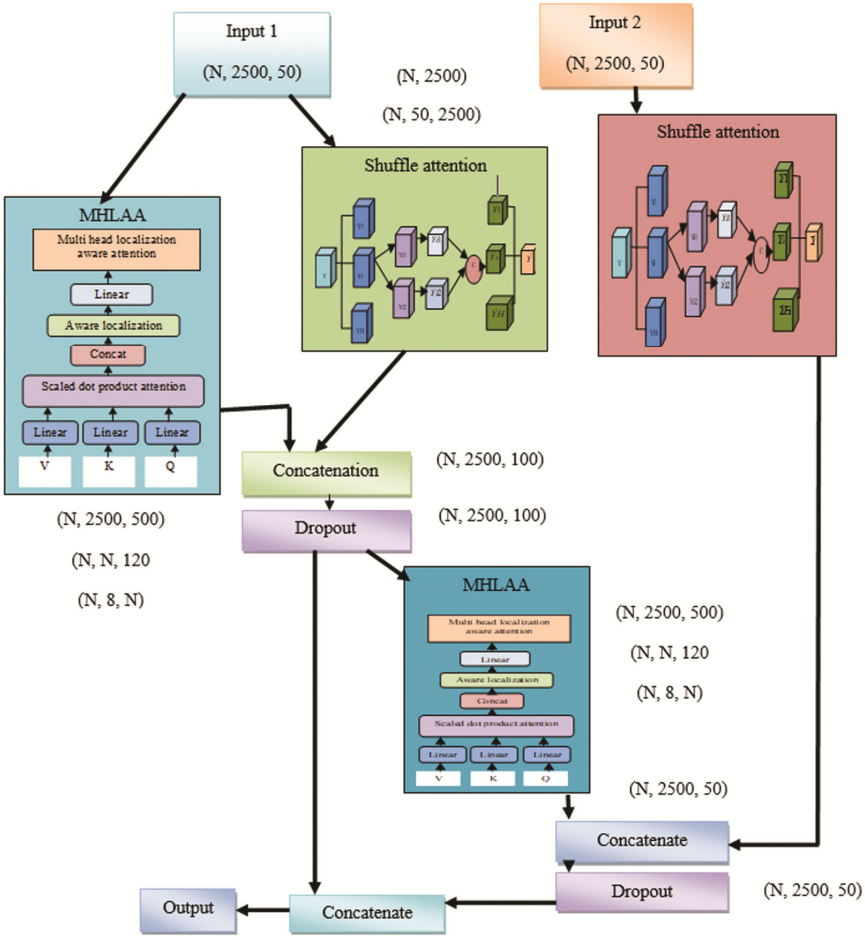

Fig. 5 — Architecture for the unified contextual shuffle attention fusion

Experimental Setup

In order to carry out the experiment for musical emotion recognition experiment, a Python program is being executed on a Windows 10 OS with 8GB of memory. *DEAM dataset* [20]*:* The dataset includes 58 songs from the 2015 evaluation set, 1000 songs from the 2014 evaluation set, and 744 songs from the 2014 development set. The development set for 2015 (provided separately), was a selection of songs from 2013 and 2014. The 45-second segments that were taken are all re-encoded with 44100Hz as the sample frequency. Additionally included in the same package are full tracks. A song's randomly (or uniformly) dispersed starting point serves as the source of the 45-second snippets. Annotations are present in 45-second segments that were taken at random from songs in the datasets from 2013 and 2014. Full song annotations are available for the dataset from 2015. The full songs as well as the 45-second excerpts are available in MPEG layer 3 (MP3) format.

Experimental Results

The experimental outcomes obtained for the MER are shown in Fig. 6. The audio signal input: Fig. 6a; the preprocessed signal: Fig. 6b; the outcome of the spectral contrast feature extraction: Fig. 6c; The MfCC feature extraction outcome: Fig. 6d; The spectral band feature extraction output: Fig. 6e; the spectral roll of feature extraction outcome: Fig. 6f; the spectral centroid feature extraction outcome: Fig. 6g, the STFT feature extraction dimension: Fig. 6h, tonal feature extraction output: Fig. 6i; VGG 16 feature extraction output: Fig. 6j; zero cross feet outcome: Fig. 6k.

Performance Analysis based on TP

The results of utilizing the proposed DualBiGRU-UCSA model for MER are shown in Fig. 7 (a–e). In Fig. 7a, the accuracy at epochs 100, 200, 300, 400, and 500 reflects percentages of 89.12%, 91.59%, 92.62%, 94.45%, and 96.28%. It is noteworthy that these outcomes consistently achieve a TP of 90. For

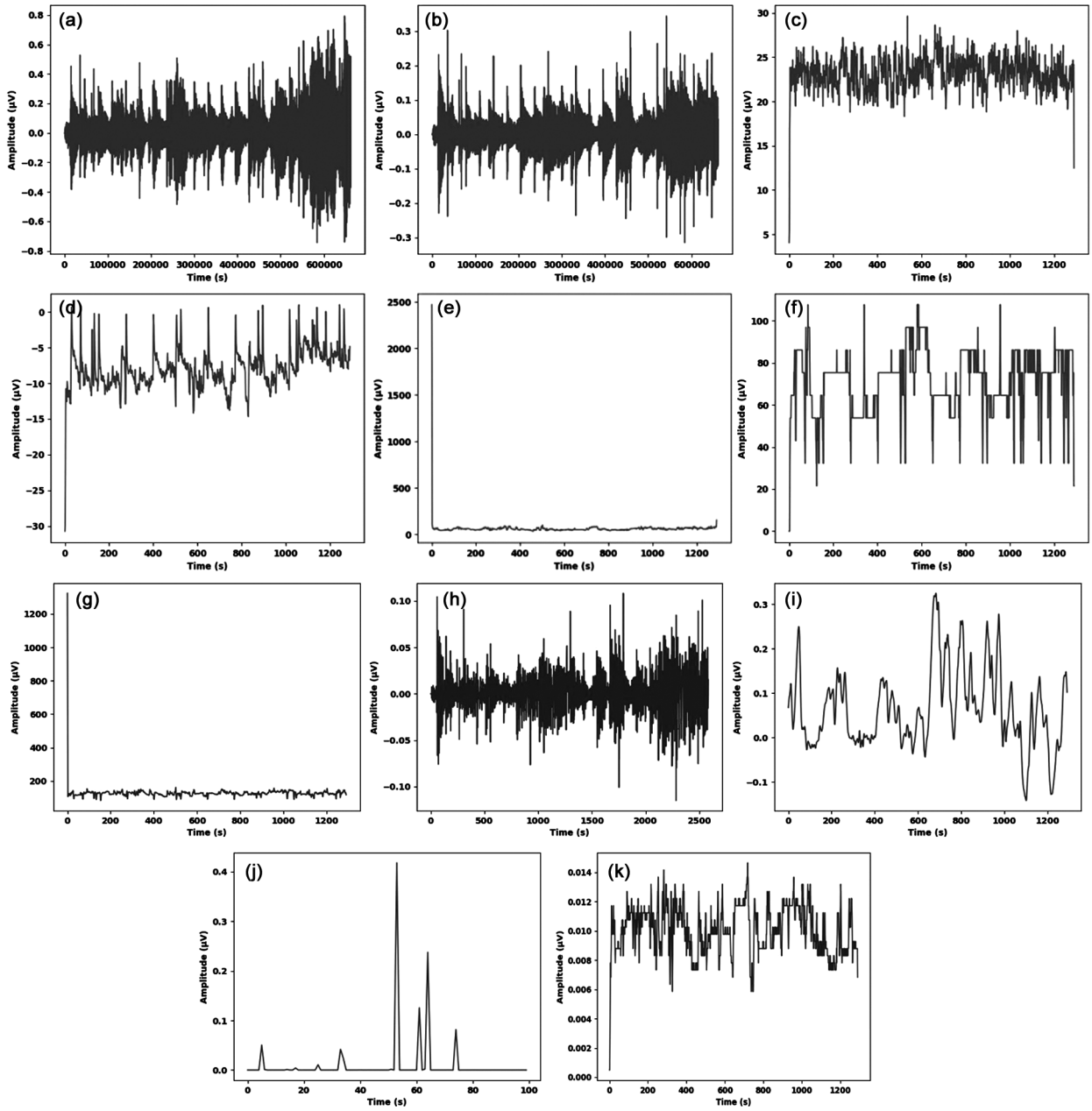

Fig. 6 — Experimental Results using the proposed DualBiGRU-UCSA model: (a) the audio signal input, (b) the preprocessed signal, (c) the outcome of the spectral contrast feature extraction, (d) The MfCC feature extraction outcome, (e) the spectral band feature extraction output, (f) the spectral roll of feature extraction outcome, (g) the spectral centroid feature extraction outcome, (h) the STFT feature extraction dimension, (i) tonal feature extraction output, (j) VGG 16 feature extraction output, (k) zero cross feet outcome

the same epoch values, the F1-score TP of 90 yields 88.76%, 92.59%, 93.49%, 94.86%, and 96.32% (Fig. 7b). Similarly Fig. 7c presents the findings of the proposed DualBiGRU-UCSA model, which is achieving NPV 89.94%, 90.45%, 91.41%, 93.78%, and 95.96% when the TP is 90. Additionally, Fig. 7d displays the results for PPV 88.30%, 92.73%, 93.83%, 95.13%, and 96.60% at TP 90. The recall rates with these same percentages of 89.22%, 92.46%, 93.14%, 94.60%, and 96.04% specifically during TP 90 shown through Fig. 7e.

Performance Analysis based on k-fold

The results from using the proposed DualBiGRU-UCSA for MER are shown in Fig. 8. The accuracy values for same epochs are 91.25%, 92.27%, 92.93%,

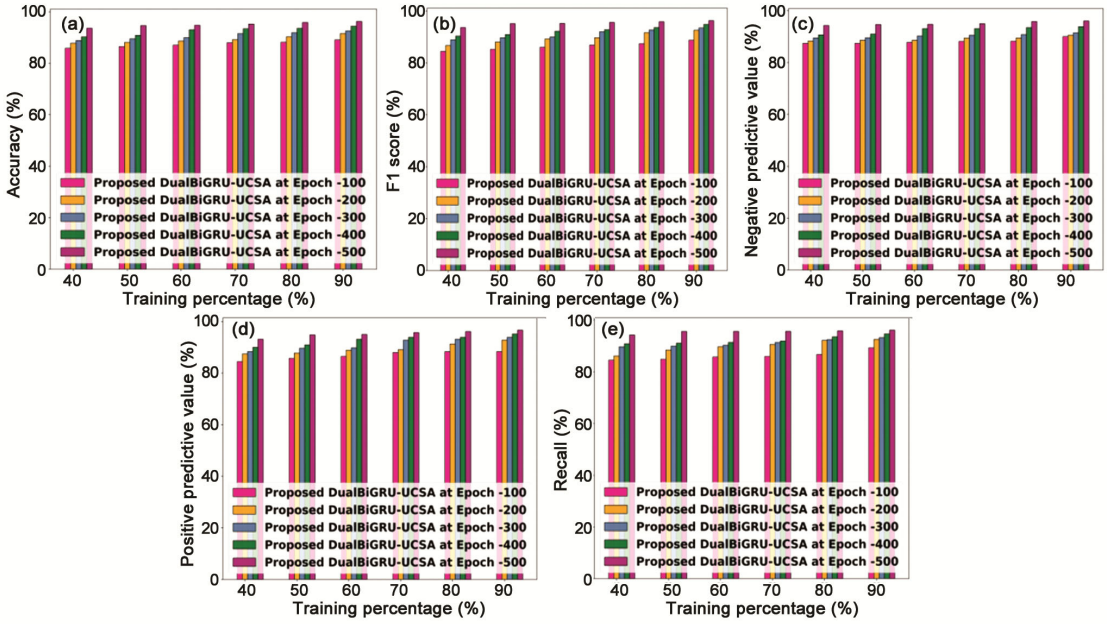

Fig. 7 — Performance Analysis based on TP (a) Accuracy, (b) F1-score, (c) NPV, (d) PPV, (e) Recall

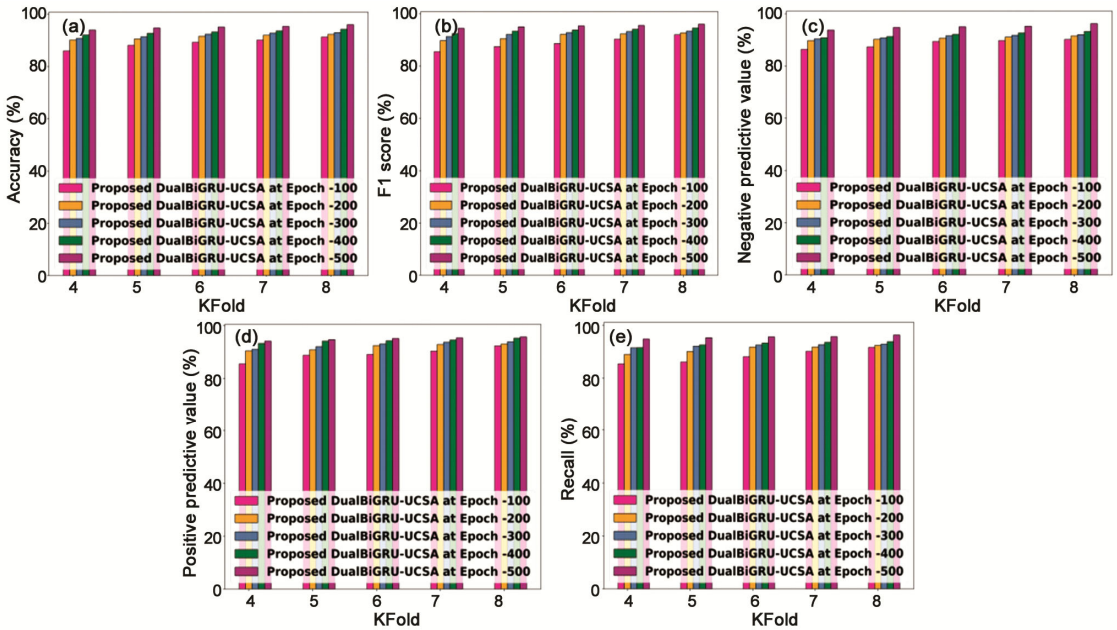

Fig. 8 — Performance Analysis based on K-fold: (a) Accuracy, (b) F1-score, (c) NPV, (d) PPV, (e) Recall

94.24%, and 95.97%, respectively (Fig. 8a). It is important to note that these results consistently reflect a k-fold of 8. For an F1-score with a k-fold of 8, the corresponding percentages for the same epochs are 91.96%, 92.67%, 93.34%, 94.51%, and 95.98% (Fig. 8b). Likewise, Fig. 8c illustrates that the DualBiGRU-UCSA model achieves NPVs of 90.18%, 91.53%, 92.01%, 93.25%, and 96.26% under the k-fold of 8. Furthermore, Fig. 8d displays the PPV results of 92.33%, 93.01%, 93.85%, 95.23%, and 95.69% at a k-fold of 8. Finally, Fig. 8e shows the recall values that align with these same percentages 91.59%, 92.33%, 92.84%, 93.80%, and 96.27% specifically for the k-fold of 8.

Comparative Methods

To illustrate the achievements of the Proposed DualBiGRU-UCSA model, a comparative analysis is

Table 1 — Comparative Discussion Table during TP 90

Models	TP 90				
	Accuracy	F1-score	NPV	PPV	Recall
Residual CNN [17]	91.48%	92.86%	91.34%	91.62%	94.13%
BLNN [6]	89.03%	86.44%	90.50%	87.56%	85.36%
Attention-based spatial-temporal model [19]	88.60%	85.96%	90.92%	86.27%	85.64%
MuSi-ABC [4]	89.00%	89.91%	89.52%	88.48%	91.38%
Embedding-based method [11]	88.79%	89.28%	88.55%	89.04%	89.53%
LSTM [12]	92.94%	94.39%	91.50%	94.39%	94.39%
DualBiGRU [21]	93.55%	94.59%	92.70%	94.39%	94.79%
Proposed DualBiGRU-UCSA	96.28%	96.32%	95.96%	96.60%	96.04%

Table 2 — Comparative Discussion Table during K-fold 8

Models	K-fold 8				
	Accuracy	F1 score	NPV	PPV	Recall
Residual CNN[17]	88.86%	88.79%	88.87%	88.86%	88.72%
BLNN[6]	89.26%	91.24%	87.36%	91.15%	91.34%
Attention-based spatial-temporal model[19]	91.17%	91.44%	89.17%	93.16%	89.79%
MuSi-ABC[4]	90.27%	88.79%	90.51%	90.02%	87.59%
Embedding-based method[11]	88.12%	87.71%	89.31%	86.93%	88.50%
LSTM[12]	92.85%	93.26%	92.28%	93.41%	93.10%
DualBiGRU[21]	95.33%	94.74%	95.98%	94.67%	94.80%
Proposed DualBiGRU-UCSA	95.97%	95.98%	96.26%	95.69%	96.27%

conducted. The analysis employs several methodologies, including as residual CNN [17], BLNN[6], attention based spatial temporal model[19], Musi-ABC[4], embedding based method[11], LSTM[12] and DualBiGRU[21] (Please refer Table 1 and Table 2).

Comparative Analysis based on TP

The proposed DualBiGRU-UCSA model with an accuracy of 96.28% outperforms the DualBiGRU model by 2.84% in the MER domain (Fig. 9a), at a TP of 90, proposed model records F1-score 96.32% (Fig. 9b), surpassing the DualBiGRU model by 1.80%, the proposed DualBiGRU-UCSA model demonstrates superior performance compared to the DualBiGRU model in MER (Fig. 9c), achieving a 3.39% enhancement and a NPV of 95.96% at a TP of 90. The proposed model excels over other models in MER (Fig. 9d) at a TP of 90, it achieves a PPV of 96.60%, indicating a 2.29% improvement over the DualBiGRU model. The proposed DualBiGRU-UCSA model also exhibits superior recall of 96.04% at a TP of 90 (Fig. 9e), again outstripping the DualBiGRU model by 1.30%.

Comparative Analysis based on K-fold

In the field of MER, the proposed model has a 0.67% performance increase compared to the DualBiGRU model, with accuracy 95.97% (Fig.10a). The proposed model surpasses the DualBiGRU model at a k-fold of 8 with F1-score 95.98%, which is 1.29% higher than that of the DualBiGRU model (Fig. 10b). Proposed model shows better results than the DualBiGRU model in MER, with a 0.29% improvement and a Negative Predictive Value (NPV) of 96.26% at a k-fold of 8 (Fig. 10c) and attaining a PPV 95.69% at a k-fold of 8, having a 1.06% enhancement over the DualBiGRU model (Fig. 10d). The proposed model achieves a higher recall of 96.27% at a k-fold of 8, exceeding the DualBiGRU model by 1.53% (Fig. 10e).

Comparative Analysis based on Time Complexity

In the number of iteration (NOI) 96 and (NOI) 98, the proposed model has lower time complexity of 18.54 and 19.13 respectively, outperforming the existing models residual CNN, BLNN, the attention-based spatial-temporal model, Musi-ABC, embedding-based methods, LSTM, and the DualBiGRU model, with time complexities of 19.26, 19.20, 19.14, 19.05, 19.03, 19.03, and 18.83, along with for (NOI) 18 time complexities of 19.50, 19.43, 19.39, 19.29, 19.27 and 19.22 respectively (Fig. 11).

Comparative Analysis based on ROC

The results of the ROC analysis for the DEAM dataset depicted through Fig. 12. The proposed model achieves an ROC score of 0.965, surpassing earlier methods residual CNN, BLNN, attention-based spatial-temporal model, Musi-ABC, embedding-based

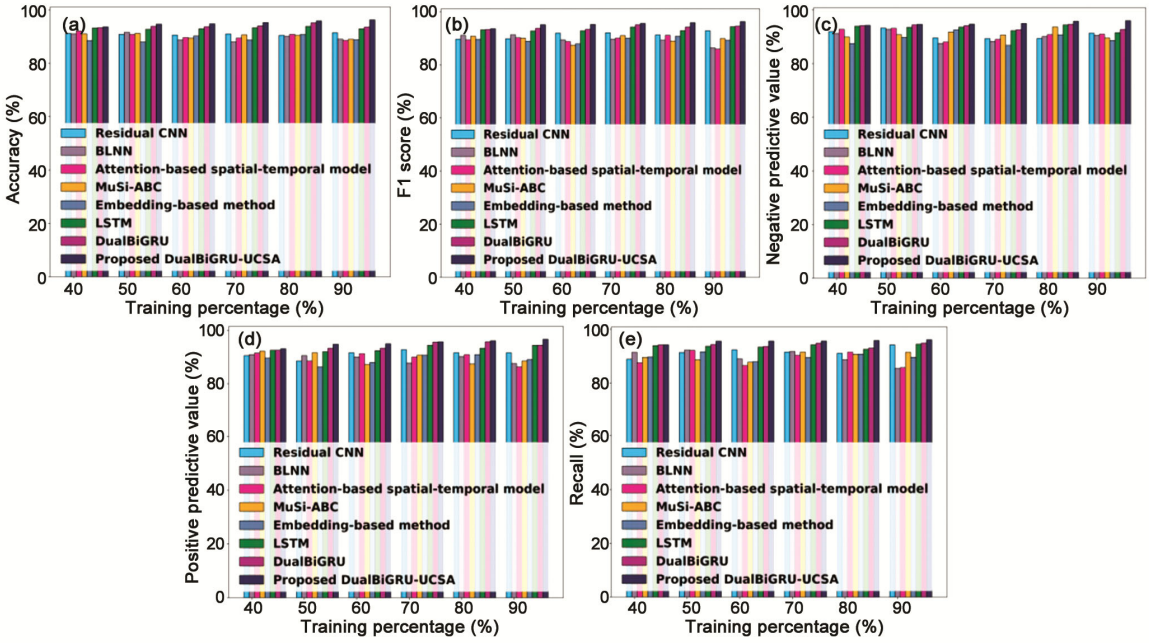

Fig. 9 — Comparative Analysis based on TP: (a) Accuracy, (b) F1-score, (c) NPV, (d) PPV, (e) Recall

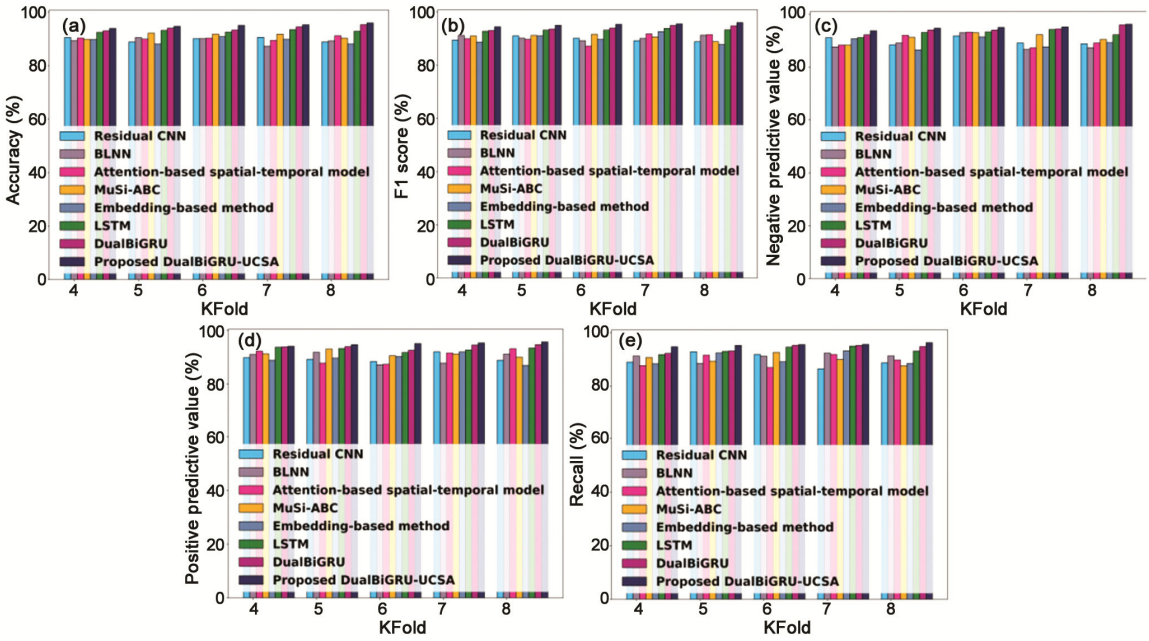

Fig. 10 — Comparative Analysis based on K-fold: (a) Accuracy, (b) F1-score, (c) NPV, (d) PPV, (e) Recall

technique, LSTM, and the DualBiGRU model, with ROC scores 0.715, 0.744, 0.752, 0.817, 0.916, 0.950, and 0.958, respectively.

Comparative Discussion

Challenges of existing Residual CNN, BLNN, and attention-based spatial-temporal models in MER includes which do not perform very well at capturing spatial and or temporal relations and this leads to discarding of valuable contextual information. Residual CNNs are very efficient in learning spatial structures of data while they are incapable of grasping temporal characteristics indispensable for the analysis of emotional content of music. Likewise, models such as BLNN and LSTM are good when it comes to dealing with sequential data yet they don't adequately

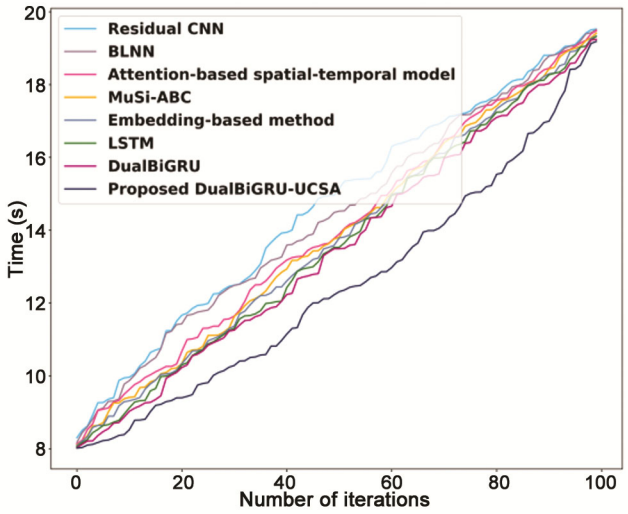

Fig. 11 — Comparative Analysis based on time complexity

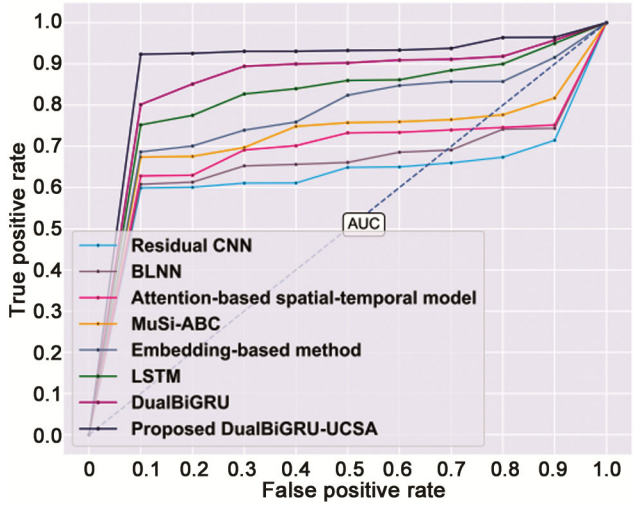

Fig. 12 — Comparative analysis based on ROC analysis

make use of spatial relations within the features fed to the model. Though attention-based models allow for better focus on informative features, they can be costly with respect to computation time and may still fail to properly capture the interactions between spatial and channel dimensions. The aforementioned challenges are addressed by the proposed DualBiGRU-UCSA in which BiLSTM is used for temporal modeling and incorporates GRU layers for enhanced feature extraction. The UCSA module again extends the model by incorporating both spatial and channel wise relations, thereby providing a stronger and contextual representation of the input data.

Conclusions

Many prior developed methods do not have complex strategies for interacting with noteworthy affective clues while considering the variability of musical characteristics. These limitations are overcome through the proposed DualBiGRU-UCSA model. It is the most effective innovation in MER which effectively combines dual bidirectional GRUs with the UCSA module to solve the problem of the complex emotional expression of music. Using bidirectional GRUs, it is possible to capture complex temporal relations within musical sequences, while UCSA further helps in improving the features' understanding due to its dynamic attention mechanisms. This synergism helps to alleviate the shortcomings of the existing approaches some of them are lack of temporal context and insufficient number of emotional nuances which results in enhanced accuracy of emotion recognition. Building upon the current use of BiLSTM for temporal modeling, future research could investigate more advanced or hybrid temporal modeling techniques. Exploring variations such as Transformer-based architectures could offer improved handling of long-range dependencies and dynamic changes in musical data. The proposed DualBiGRU-UCSA model attains high accuracy, f1-score, NPV, PPV and recall of 96.28%, 96.32%, 96.26%, 96.60% and 96.27% respectively.

References

1 Doyran M, Schimmel A, Baki P, Ergin K, Türkmen B, Salah A A, Bakkes S C, Kaya H, Poppe R & Salah A A, MUMBAI: multi-person, multimodal board game affect and interaction analysis dataset, *J Multimodal User Interfaces*, **15(4)** (2021) 373–391.

2 Camacho M C, Williams E M, Ding K & Perlman S B, Multimodal examination of emotion processing systems associated with negative affectivity across early childhood, *Dev Cogn Neurosci*, **48** (2021) 100917.

3 Liu D, Chen L, Wang Z & Diao G, Speech expression multimodal emotion recognition based on deep belief network, *J Grid Comput*, **19(2)** (2021) 22.

4 Yang J, musi-ABC for Predicting Musical Emotions, *IEEE Access*, **11** (2023) 79455–79465, doi: 10.1109/ACCESS.2023.3300042.

5 Ballantine C, Against populism: music, classification, genre, *Twent-Century Music*, **17(2)** (2020) 247–267, https://doi.org/10.1007/s00500-024-09922-6.

6 Du X, BLNN: A muscular and tall architecture for emotion prediction in music, *Soft Comput* **28(20)** (2024) 11855–11871.

7 Huang C & Zhang Q, Research on music emotion recognition model of deep learning based on musical stage effect, *Sci Program*, **1** (2021) 3807666.

8 He N & Ferguson S, Music emotion recognition based on segment-level two-stage learning, *Int J Multimed Inf Retr*, **11(3)** (2022) 383–394.

9 Jaiswal M & Provost E M, Privacy enhanced multimodal neural representations for emotion recognition, *In Proc of the AAAI Conf on Artificial Intell*, **34(05)** (2020) 7985–7993.

10 Choi D Y, Kim D H & Song B C, Multimodal attention network for continuous-time emotion recognition using video and EEG signals, *IEEE Access*, **8** (2020) 203814–26.

11 Takashima N, Li F, Grzegorzek M & Shirahama K, Embedding-based music emotion recognition using composite loss, *IEEE Access*, **11** (2023) 36579–604.

12 Chen W, A novel long short-term memory network model for multimodal music emotion analysis in affective computing, *J Appl Sci Eng*, **26(3)** (2022) 367–376.

13 Xia K, Hu T & Si W, Editorial for the special issue on Research on methods of multimodal information fusion in emotion recognition, *Pers Ubiquitous Comput*, **23** (2019) 359–361.

14 Malik M, Adavanne S, Drossos K, Virtanen T, Ticha D & Jarina R, Stacked convolutional and recurrent neural networks for music emotion recognition, *arXiv preprint arXiv,* (2017) 1706.02292.

15 Zhao Z, Bao Z, Zhao Y, Zhang Z, Cummins N, Ren Z & Schuller B, Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition, *IEEE Access*, **7** (2019) 97515–97525.

16 Gorrostieta C, Brutti R, Taylor K, Shapiro A, Moran J, Azarbayejani A & Kane J, Attention-based Sequence Classification for Affect Detection, *In Interspeech*, (2018) 506–510.

17 Han X, Chen F & Ban J, Music emotion recognition based on a neural network with an Inception-GRU residual structure, *Electronics*, **12(4)** (2023) 978.

18 Raboy L J & Taparugssanagorn A, Verse1-Chorus-Verse2 Structure: A stacked ensemble approach for enhanced music emotion recognition, *Appl Sci*, **14(13)** (2024) 5761.

19 Su Y, Chen J, Chai R, Wu X & Zhang Y, FFA-BiGRU: Attention-Based Spatial-Temporal feature extraction model for music emotion classification, *Appl Sci*, **14(16)** (2024) 6866.

20 DEAM dataset : https://cvml.unige.ch/databases/DEAM/ (Accessed on: 25 Feb 2024)

21 Wang Y, Feng L, Liu A, Wang W & Hou Y, Dual BiGRU-CNN-based sentiment classification method combining global and local attention, *J Supercomput*, **80(2)** (2024) 2799–2837.

22 Ye W, Wang J, Chen L, Dai L, Sun Z & Liang Z, Adaptive spatial–temporal aware graph learning for EEG-based emotion recognition, *Cyborg Bionic Syst*, **5** (2024) 0088, https://doi.org/10.34133/cbsystems.0088.

23 Pan S, Xu G J W, Guo K, Park S H & Ding H, Cultural insights in souls-like games: analyzing player behaviors, perspectives, and emotions across a multicultural context, *IEEE Trans Games*, **16(4)** (2024) 758–769, doi: 10.1109/TG.2024.3366239.

24 Zhu C, Research on emotion recognition-based smart assistant system: emotional intelligence and personalized services, *J Syst Manag Sci*, **13(5)** (2023) 227–242, doi: 10.33168/JSMS.2023.0515.

25 Ding J, Chen X, Lu P, Yang Z, Li X, Du Y, DialogueINAB: An interaction neural network based on attitudes and behaviours of interlocutors for dialogue emotion recognition, *J Supercomput*, **79(18)** (2023) 20481–20514, doi: 10.1007/s11227-023-05439-1.

26 Song L, Chen S, Meng Z, Sun M and Shang X, FMSA-SC: A fine-grained multimodal sentiment analysis dataset based on stock comment videos, *IEEE Trans Multimed*, **26** (2024) 7294–7306, doi: 10.1109/TMM.2024.3363641.