Music Emotion Recognition

Chandrasena M.M.D.
Department of Computer Engineering
University of Peradeniya
Sri Lanka
e17040@eng.pdn.ac.lk

Upekha H.P.S.
Department of Computer Engineering
University of Peradeniya
Sri Lanka
e17356@eng.pdn.ac.lk

Wijesooriya H.D.
Department of Computer Engineering
University of Peradeniya
Sri Lanka
e17407@eng.pdn.ac.lk

Prof. Roshan Ragel
Department of Computer Engineering
University of Peradeniya
Sri Lanka
roshanr@eng.pdn.ac.lk

Dhanushki Mapitigama
MSc Student in Data Science
Uppsala University
Sweeden
dhanumapitigama@gmail.com

***Abstract*—Music can be considered as a universal language of emotions. Music Emotion Recognition (MER) is an area of research that focuses on the algorithms and techniques to recognize and understand those emotions, which is mainly used in personalized music recommendation systems, music therapy, and effective computing fields. In this paper, we propose an approach for dynamic music emotion recognition based on bidirectional long short-term memory (BiLSTM) deep neural network architecture along with multiple dense layers. Two models were trained separately for the arousal and valence. To evaluate our model, we used the 'DEAM' dataset which is the most widely used dataset in the research field of MER. In our methodology, we used principal component analysis (PCA) in the feature selection part and then those data were sent to the BiLSTM model. Next, the output of the BiLSTM model was sent through multiple dense layers. Finally, the emotion that is expressed by the given music sample was determined by the model output and Thayer's model. The results show that our model outperforms the traditional machine learning algorithms like linear regression (LR), support vector regression (SVR), etc, and our reference model which is the BiLSTM model.**

***Index Terms*—Music Emotion Recognition, MER, Bidirectional Long Short-Term Memory, BiLSTM, feature extraction, Principal Component Analysis, PCA, machine learning algorithms, Linear Regression, LR, Support Vector Regression, SVR, Deep Neural Network**

I. Introduction

Music has a great ability to generate various emotions like happiness, sadness, excitement, anger, and many more emotions in the listener's mind. When we consider Music emotion recognition, it is an example of a field that uses machine learning and neural network techniques. Those machine learning algorithms are trained by emotional characteristics that are included in the music compositions like melody, tempo, rhythm, harmony, timbre, etc.

There are several advantages of MER systems like in the field of music therapy where therapists can choose appropriate music that is in line with their client's emotional needs. Also for the creation of personalized music recommendation systems and playlists based on listeners' emotional preferences. In addition to that, MER systems are very useful in the media and entertainment sector since they can improve the audiovisual experience by analyzing the emotional content of the music.

During the literature survey, we went through the methods and technologies used by the previous researchers and identified their drawbacks. So, we found that the accuracies of the existing systems are much lower. Therefore the main objective of this research is to develop a more accurate dynamic MER system.

When developing dynamic MER systems, most of the researchers have used LSTM and BiLSTM models in those systems [[7], [9], [10], [12], [13]]. Those researchers have proven that the deep learning models are more accurate than the traditional machine learning algorithms. In the paper [22], they have used the concept of DNN (Dense Neural Networks) to improve the accuracy of their system. So, in this research, we came up with a methodology that contains the BiLSTM model and the DNN concept.

II. Literature Review

A. Preliminary Knowledge

Existing MER publications can be divided into two sections, namely song-level MER (or static) and music emotion variation detection (MEVD, or dynamic). Assigning the overall emotion label to one song is known as song-level MER. MEVD considers the emotion of the music as a changing process.

- **Research framework:**
  MER systems contain three main parts and they are, domain definition, feature extraction, and emotion recognition. "Fig. 1", shows the overall framework of MER systems. According to the MER framework, initially, emotion models and datasets are selected in the domain

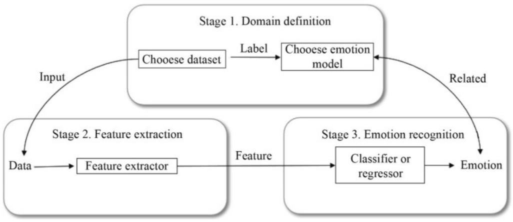

Fig. 1. Music Emotion Recognition framework

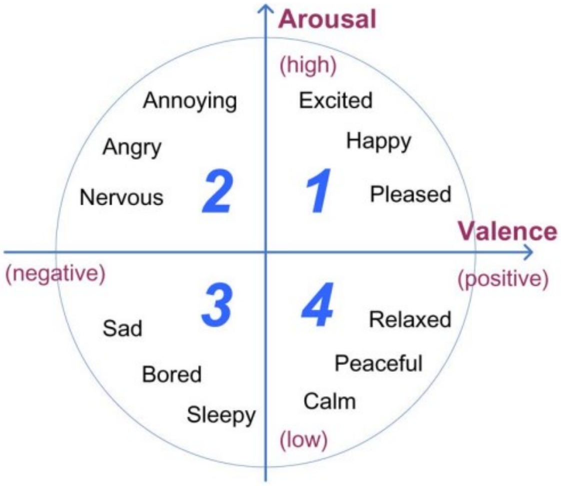

Fig. 2. Thayer's arousal-valence emotion plane

definition stage. Useful features are extracted in the feature extraction stage, and after that, the emotion label is predicted in the emotion recognition stage.

- **Emotion Model:**
  "Table. I" summarizes widely used emotion models in MER. In the "Emotion Conceptualization" column, "Categorical" refers to the categorical emotion model, and "Dimensional" means the dimensional emotion model.
  Dimensional emotion models are widely used in MER systems. There are two main dimension emotion models namely Thayer's emotion model and Russell's circumplex model. Both models use arousal and valence values (AV values) to identify the emotion in a given music sample."Fig. 2" shows Thayer's emotion model associated with MER.
- **Datasets:**
  "Table. II" lists some information about the most commonly used public datasets.

B. Related Works

The main objective of this review is to get an idea about the current state of machine-learning-based music emotion recognition systems. The Keywords such as Music Emotion Recognition, Music Mood, and Speech Emotion Recognition were selected at the beginning. After that, papers from Google Scholar, and ResearchGate databases were searched up to 2022. Finally, we went through the papers that were published in recent years.

TABLE I
A SUMMARY OF EMOTION MODELS

Model name	Emotion conceptualization	Number of classes /dimensions
Hevner affective ring	Categorical	67
Russell's model	Dimensional	2
Thayer's model	Dimensional	2

TABLE II
A SUMMARY OF DATASETS

Dataset name	Emotion conceptualization	Number of songs	Research directions
MediaEval	Dimensional	100	Dynamic
CAL500	Categorical	500	Static
AMG1608	Dimensional	1608	Static
DEAM	Dimensional	1802	Dynamic
PMEmo	Dimensional	1000	Dynamic

In the research [9], "DBLSTM-Based Multi-Scale Fusion for Dynamic Emotion Prediction in Music" by Xinxing Li, Jiashen Tian, Mingxing Xu, Yishuang Ning, and Lianhong Cai, proposed a regression-based method to predict the continuous emotion change in music. The researchers have used the emotion model proposed by Russel and the MediaEval 2015 dataset. The emotion in music is associated with both previous and future content. There is the ability to use both previous and future information, in Bidirectional Long Short-Term Memory (BLSTM). The LSTM model is good at exploiting and storing information for long time periods. BLSTM is developed based on LSTM, therefore it is capable of exploiting context for long periods of time while reaching the context in both previous and future directions.

First, they have input features to multiple DBSTLM models with different sequence lengths to predict AV values. Here, the BLSTM models were trained for valence and arousal separately. After that post-processing and fusion components were applied to each individual output of the DBSTLM models. Here, they have applied post-processing to the individual output of DBLSTM, to make use of temporal correlation in music. The fusion component was used to integrate the outputs of all DBSTLM models with different scales. Here they tried different orders of applying post-processing and fusion components and in the end, they found that the Post-processing after fusion gave the best result. To find the best network structure, they compared the performance of BLSTM models with different numbers of layers and units on the validation data. Finally, the BLSTM models with 5 layers and 250 units were used.

In the study “A Deep Bidirectional Long Short- Term Memory Based Multi-scale Approach for Music Dynamic Emotion Prediction ” by Xinxing Li, Haishu Xianyu, Jiashen Tian, Wenxiao Chen, Fanhang Meng, Mingxing Xu, and Lianhong Cai [10], proposed a Deep BLSTM (DBLSTM) based multi-scale regression and fusion with Extreme Learning Machine (ELM), to predict the valence and arousal values in music. MediaEval - 2015 dataset was used in this research.

First, they cut the complete songs into different sequence lengths: 10s, 20s, 30s, and 60s. Then, those data were input into the four kinds of DBLSTM models, and those were trained with different sequence scales of 10, 20, 30, and 60, respectively. Here, they have trained the DBLSTM models separately for arousal and valance. Finally, the outputs of DBLSTM models have been input to ELM, and ELMs were trained for valence and arousal separately. Extreme Learning Machine (ELM) is a learning algorithm for single-hidden layer feedforward neural networks (SLFN).In order to find the best network structure, the performance of DBLSTM models were compared with different number of layers and units on the validation data. In this study, they have compared the accuracy of the LSTM model with other regression models such as SVR, MLR, etc and it can be seen that the LSTM model has the highest accuracy.

In the research study [7], ”Multi-scale Context Based Attention for Dynamic Music Emotion Prediction” by Ye Ma, XinXing Li, Mingxing Xu, Jia Jia, and, Lianhong Cai, developed a system to recognize the continuous emotion information in music. A two-dimensional valence-arousal emotion model was used to represent the dynamic emotion music. The proposed method was evaluated using the MediaEval 2015 dataset. The proposed method contains a Long Short-Term Memory (LSTM) based sequence-to-one mapping. By using this sequence-to-one music emotion mapping, they have proved the influence of different time scales’ preceding content on the LSTM model’s performance. Therefore they further proposed a multi-scale Context-based Attention (MCA) mechanism. This mechanism was used to give different time scales’ preceding context respective attention weights, as the music emotion at a specific time is the accumulation of a piece of music content before that time point. As it is difficult to determine how much previous content is suitable for the emotion prediction, they paid different attention to the previous context of different time scales, and the weights of different scales were dynamically computed by the model.

First, feature sequences with different lengths were input into the LSTM models. Then, each individual output of LSTM models was sent through a context vector. Next, that output was sent to the MCA model. Finally, they predicted the valence and arousal values based on the weighted sum of the multi-scale context vector, which was output by the MCA model. To demonstrate the effectiveness of the MCA model, they have done three sets of experiments, using single-scale LSTM, attention-based LSTM, uniform MCA, and MCA model. According to their results, it can be seen that the attention-based LSTM with MCA has the highest accuracy compared to the other combinations. And also, compared the accuracies of their method with other models such as MLR, SVR, LSTM, etc, and proved that the proposed method outperformed them.

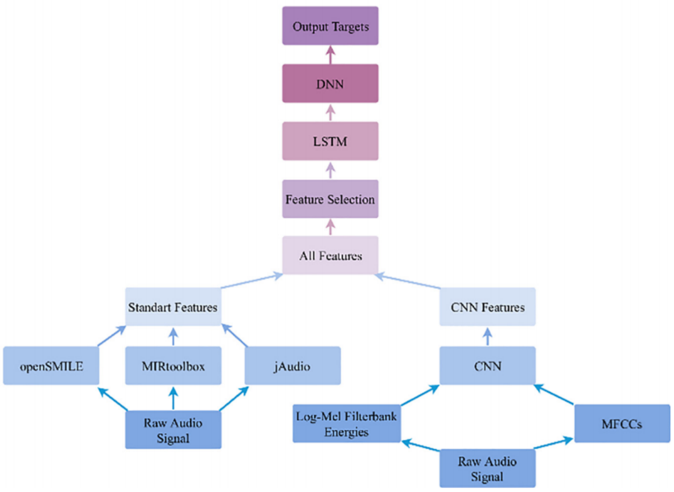

Fig. 3. The architectures of convolutional long short-term memory deep neural network.

The paper ”Music Emotion Recognition Using Convolutional Long Short Term Memory Deep Neural Networks,” [22] suggests a method for MER using convolutional long short-term memory deep neural network (CLDNN) architecture design. They have used a new Turkish emotional music database with 124 different Turkish traditional songs each of length 30 seconds. Here log-mel filterbank energies and mel frequency cepstral coefficients (MFCCs) are utilized as features to achieve high performance. Also, this paper emphasizes the challenges associated with labeling emotions, feature extraction, and selection of suitable classification algorithms. A dimensional model with arousal and valance for emotion annotations is used to address those challenges. That explores a different range of acoustic features relevant to music and emotion. This covers aspects such as pitch, melody, harmony, tonality, timing, dynamics, and rhythm. Their proposed CLDNN architecture combined CNN for feature extraction and LSTM + DNN for classification. This outperforms other classifiers like k-nearest neighbor (k-NN), support vector machine (SVM), and Random Forest, achieving a higher overall accuracy through 10-fold cross-validation.“Fig. 3” represents the overall procedure followed in this research paper.

The paper, ”Regression-Based Music Emotion Prediction Using Triplet Neural Networks” by Kin Wai Cheuk, Yin-Jyun Luo, Balamurali B.T, Gemma Roig, and Dorien Herremans [2] used a regression-based approach to predict the emotion in music. There, they used triplet neural networks (TNN) to perform a regression task and the predictions were done according to the arousal and valence values.TNNs were initially introduced for classification but here it was used to provide low-dimensional representation for regression task. That is, they used TNN as a dimensionality reduction method. Both the DEAM dataset and the 2013 MediaEval dataset were used.

They implemented their novel TNN regression approach for dimensionality reduction and combined it with both a support vector regressor (SVR) and a gradient boosting machine (GBM) to solve the regression problem for the valence and arousal values. They first tested the system using the 2013 - MediaEval dataset, which was called as ”MediaEval experiment”. The TNN implemented in this experiment contains a single fully connected layer with 600 neurons and ReLU as the activation function. Here they compared the TNN results against other dimensionality reduction methods such as principal component analysis (PCA), Gaussian random projection (RP), and neural network-based autoencoder (AE). From this experiment, it was found that the TNN-based models performed best when fewer layers were used. Therefore, they used a single-layer fully connected network with ReLU activation as their TNN structure. After that, they tested the system using the DEAM dataset which was named the ”DEAM experiment”. At the end of this research, it was found that their TNN method outperforms the widely used dimensionality reduction methods such as principal component analysis (PCA) and autoencoders (AE).

In the research paper ”Music Emotion Recognition based on Segment-Level Two-Stage Learning” [12], Na He and San Ferguson present an innovative two-part framework. The first part focuses on unsupervised learning, skillfully generating feature representations for segment-level music without the need for emotion labels. Extracting meaningful representations from music segments is the primary aim.

The second part adopts a supervised training model, treating segments as sequential units within each music excerpt. Deep learning techniques tailored for time-series data are employed to predict the final music emotion.

The paper incorporates a Convolutional Neural Network (CNN) module, reusing the feature encoder’s structure from unsupervised learning. A Bidirectional Long Short-Term Memory (BiLSTM) is used for emotion classification, with both the CNN and BiLSTM jointly trained during the supervised learning stage. This cohesive approach enables effective emotion predictions.

Insights from the PMEmo dataset reveal that 1-second segments achieve the best valence results (accuracy: 79.01%, F1-score: 83.2%), while 5-second and 10-second segments exhibit higher accuracy (83.62%/83.51%) and F1-scores (86.52%/86.62%) for arousal across the All Music dataset, emphasizing segment duration’s impact on emotion recognition.

III. Methodology

Raw audio signals were used as the primary input for the proposed music-emotional recognition system. Opensmile Python library is a powerful library for feature extraction. Opensmile python module helps to process these unprocessed audio signals and finally, usable information was given.

Feature selection is done second after extracting relevant features from the OpenSMILE library. At this point, two different feature selection methods were used as follows, Correlation-based Feature Selection (CFS) and Principal Component Analysis (PCA) Both of them were checked to enhance the effectiveness and efficiency of the system. Then we chose the PCA as our feature selection method. Subsets of features that demonstrated high relevance to target variables(arousal and valence) were systematically identified by CFC and redundancy was significantly reduced. At the same time, PCA was used to transform the original feature space into a lower dimensional representation and detect maximum variance in data.

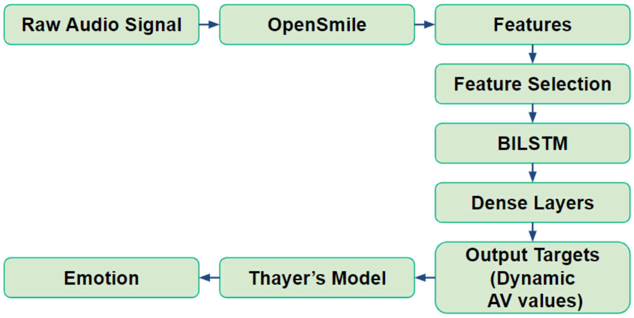

Fig. 4. Methodology

Our music emotion identification model was trained using hybrid architecture following feature selection. Long Short-Term Memory (BiLSTM) networks that are bidirectional were selected. because of their capacity to record the temporal relationships between consecutive data points. To improve the model’s ability to identify complex patterns in the feature space, strategic dense layer placement was added to the BiLSTM layers. Part of the training process included hyperparameter modification and repeated epochs. Because of the synthesis between BiLSTM and dense layers, our model was able to learn and articulate correlations between returned features and emotional states.

In the ultimate phase of our methodology, we utilized Thayer’s arousal-valence model to discern and assign specific emotions to each music segment. It is a theoretical model to recognize the emotional states. Thayer’s model is rooted in the dimensions of arousal and valence axes. we derived distinct emotional labels for each music segment, By mapping the arousal and valence predictions obtained from our machine learning model onto this established model. Then we can translate the quantitative arousal and valence values into qualitative emotional descriptors. The comprehensive integration of feature extraction, selection, machine learning with BiLSTM and dense layers, and emotion mapping through Thayer’s model collectively constituted a systematic approach for identifying and characterizing emotions in musical segments.

“Fig. 4” shows the overall procedure of the proposed system.

IV. Experiments

A. Dataset Description

The dataset that we have used to validate our model is the DEAM dataset which is also known as the ”MediaEval Database for Emotional Analysis in Music”.This dataset is publicly available and we obtained it from ”Kaggel”.This is the most widely used dataset in the research field of MER [[2], [7], [8], [9], [10], [11]]. There are 1802 songs and those are annotated with both arousal and valence values for every 0.5s. There are 260 standard audio features and those have been extracted by the openSMILE tool. There are two target variables which are known as ’Arousal’ and ’Valence’.Those target variables contain numerical values and two models have been trained separately to do the predictions.

B. Performance of Different Models

Using a machine learning methodology, a baseline model was developed for music emotion recognition. The initial algorithm we used was linear regression. The DEAM dataset was used, because it’s a well-known and extensive dataset in the field of music emotion recognition. A solid basis for our investigations is provided, with a wide variety of musical snippets tagged with emotional classifications. We used a 5-fold cross-validation technique as part of our cross-validation plan. The dataset was divided into five subsets for this purpose. Among them, four of the subsets were used to train the model, and one subset was used for validation throughout each iteration. The validation set was rotated across the folds to sufficiently evaluate the generalization capabilities. To determine how well the baseline linear regression model captured the complex relationships between musical characteristics and emotional states, we carried out a thorough review procedure.

Within the context of music emotion recognition, potential improvements are offered by support vector regression (SVR). It is established as a baseline linear regression model. For training and evaluating the SVR model, the same DEAM dataset was used. The SVR model can effectively capture complex nonlinear relationships. To ensure a robust evaluation process 5 fold cross-validation method is used here also. Here, the dataset was divided into five subsets, and among them, four subsets were used to train and one was used for validation in each fold. It allows us to investigate the impact of nonlinearity in the modeling of emotional features by SVR.

After that according to the relevant research papers that we referred to, we integrated a Bidirectional Long Short-Term Memory (BiLSTM) model in our research also. Then our model is better able to identify emotions that have complex patterns and temporal correlations in the music data. Here also we used a 5-fold cross-validation configuration. The rectified linear unit (ReLU) was used as the activation function, a learning rate of 0.001, 25 epochs, a batch size of 32, and the Adam optimizer were the hyperparameters that were carefully adjusted to maximize the model’s performance. Based on their proven efficacy in comparable trials, these criteria were chosen.

Further optimizing our music emotion recognition system, we introduced an innovative model by combining Bidirectional Long Short-Term Memory (BiLSTM) with a Deep Neural Network (DNN). With the help of this cooperative design, the system’s overall performance should be enhanced by utilizing the feature extraction powers of the DNN and the temporal relationships recorded by the BiLSTM. We continued to use a 5-fold cross-validation process as part of our commitment to thorough review. Hyperparameter settings are a learning rate of 0.001, 25 training epochs, a batch size of 32, and the Adam optimizer. They were carefully tuned to align with successful configurations in our prior models. A summary of the models we tried and the initial model parameters of them are shown in the“Table. III”

TABLE III
Summary of models

Model	Parameters
Linear Regression (baseline model)	Folds = 5
SVR	Folds = 5
BiLSTM (reference model)	Folds = 5, Learning Rate=0.001, Epochs=25, Batch Size=32, Optimizer=Adam, Activation Function=relu
BiLSTM + DNN (proposed model)	Folds = 5, Learning Rate=0.001, Epochs=25, Batch Size=32, Optimizer=Adam, Activation Function=relu, Dense Layer Units=512

C. Feature Selection

Under the feature selection, we used two distinct feature selection methods. They are Correlation-based Feature Selection (CFS) and Principal Component Analysis (PCA). The CFS method systematically identifies the subsets of features that exhibit high relevance to the target variable while minimizing redundancy and facilitating a more compact and informative feature set. PCA can transform the original feature space into a lower-dimensional representation. And it can capture the maximum variance in the data. To visually recognize the impact of these feature selection methods, graphs were plotted that show the changing landscape of feature importance and variance explained. To increase the model efficiency and overall performance the feature selection process is needed.

D. Parameter Tuning

In the parameter tuning phase of this research, we conducted some experiments to identify the optimum learning rate and the ideal number of epochs for the arousal and valence models separately. The learning rate is a hyperparameter, which determines the step size at each iteration while moving toward a minimum of a loss function. An epoch is one complete pass through the entire training dataset during the training of a model. It is very important to find the optimum number of epochs as too few epochs may result in underfitting, where the model hasn’t learned the underlying patterns in the data. On the other hand, too many epochs may lead to overfitting, where the model becomes too specific to the training data and performs poorly on unseen data.

When finding the optimum learning rate, first we changed the learning rate of the model and obtained the corresponding MSE of the model. Then we plotted a graph with those values and selected the learning rate with the lowest MSE as the optimum learning rate.

When finding the optimum number of epochs, we first changed the number of epochs and obtained the corresponding train and test set accuracies. Finally, we plotted a graph, and using this graph we selected the optimum number of epochs.

E. Application of Dense Layers

Further, to improve the prediction ability of our model we used the concept of dense neural networks. So, under this section, we tried different dense layer combinations for both arousal and valence models separately. When conducting the experiment, we changed the number of dense layers and their units to obtain the corresponding accuracy, R2 score, and MSE values. The different dense layer combinations we tried for the models are shown in the “Table. IV”.

TABLE IV
DIFFERENT DENSE LAYER COMBINATIONS

No. of Layers	No. of Units
1	512
2	512, 256
3	512, 256, 128
4	512, 256, 128, 64
5	512, 256, 128, 64, 32
5	1024, 1024, 1024, 1024, 1024

F. Model Setup

According to the results that were obtained during the above experiments, we have finalized the model parameters of the arousal and valence models. The ”TensorFlow” framework was used to implement those models. All those experiments were done on the ”T4 GPUs” available in the Google Colab free edition due to their high performance. Since we used GPUs to run the models, we had to use CuDNNLSTM instead of LSTM to construct our BiLSTM model. Because, CuDNNLSTM is specifically created for CUDA parallel tasks and requires a GPU to function, while LSTM is designed for regular CPUs. The model parameters and their values used in the final models are given in the “Table. V”.

TABLE V
FINAL MODEL PARAMETERS

Parameter	Arousal Model	Valence Model
No. of Folds	5	5
Learning Rate	0.001	0.001
Epochs	55	60
Batch Size	32	32
Optimizer	Adam	Adam
Activation Function	relu	relu
Dense Layer Units	512,256,128,64,32	512,256,128,64,32

V. RESULTS AND DISCUSSION

A. Dataset

After analyzing the dataset we found that there is an almost symmetrical distribution of the arousal and valance values of the data points in the data, the “Fig. 5” shows the distribution of arousal values, and the “Fig. 6” shows the distribution of valence values respectively.

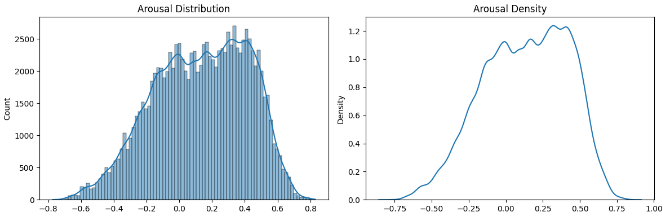

Fig. 5. Arousal Distribution

In arousal distribution, a skewness value of -0.2738 was obtained. That indicates that the data distribution is approximately symmetrical but there was a slight leftward tail. Similarly, for the valence distribution, a skewness value of -0.2319 was obtained. Although the majority of the values are concentrated towards the right side of the distribution, there is still some degree of balance since those values are closer to zero.

B. Initial Performance Comparison of Different Models

“Fig. 7” shows the accuracy, R2 score, and the RMSE of the implemented baseline model (Linear Regression), SVR model, Reference Model, and our implemented model for the Arousal model. There, we found that models that use deep neural network concepts have more accuracy compared to the models using traditional machine learning algorithms. Also, we found that the accuracy of the BILSTM model can be increased by adding more dense layers to it. At the initial stage of the implementation, we could achieve an improvement of 15.08% in R2 score, and 6.04% in accuracy and could reduce the RMSE by 10.38% when compared to the baseline model which is LR.

Similarly, “Fig. 8” shows the accuracy, R2 score, and the RMSE of the implemented baseline model (Linear Regres-

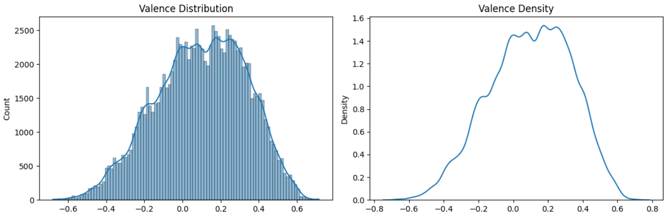

Fig. 6. Valence Distribution

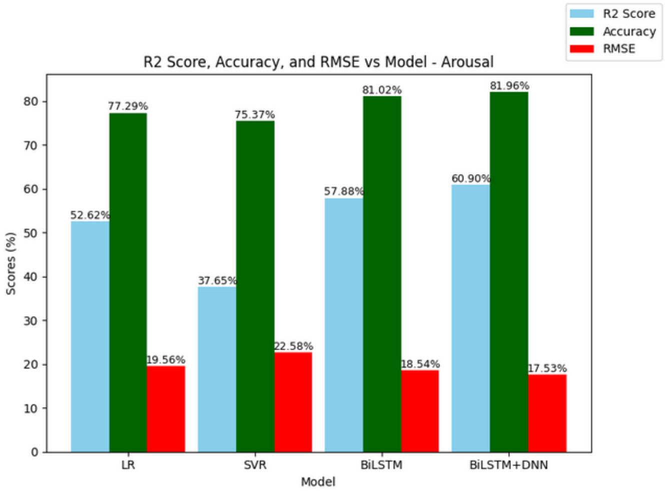

Fig. 7. Initial Performance Comparison of Arousal models

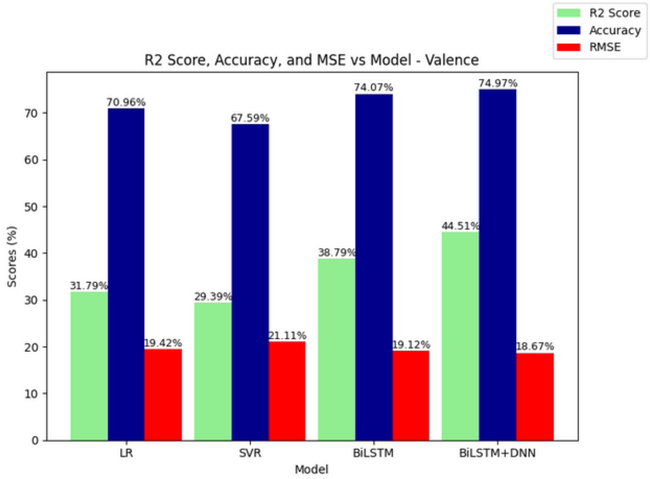

Fig. 8. Initial Performance Comparison of Valence models

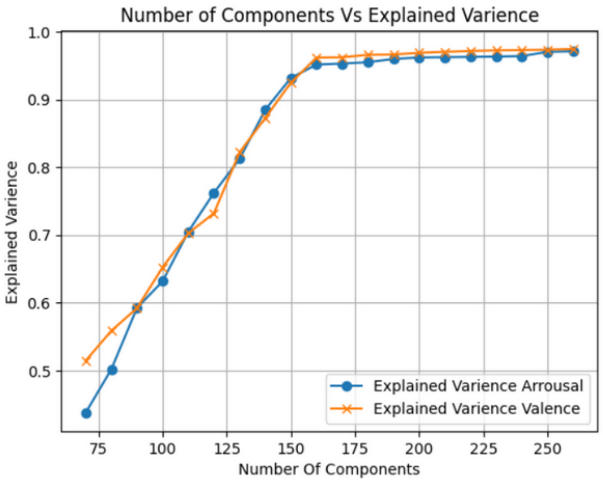

Fig. 9. Number of components Vs Explained Varience

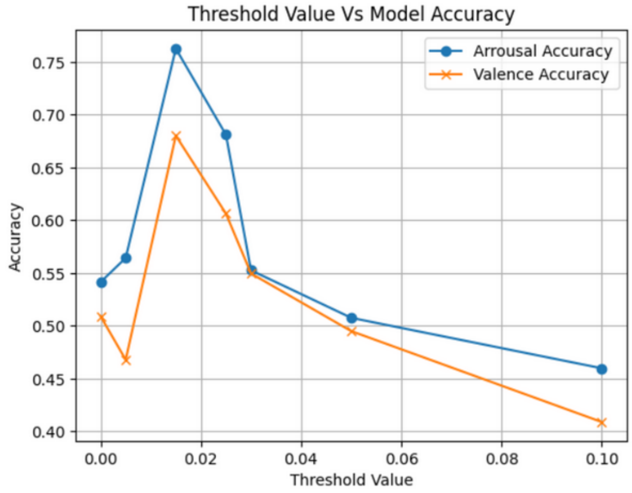

Fig. 10. Threshold Value and the Model Accuracy

sion), SVR model, Reference Model, and our implemented model for the valence model. At the initial stage of the implementation of the valence model, we could achieve an improvement of 40.01% in R2 score, and 5.65% in accuracy and could reduce the RMSE by 3.86% when compared to the baseline model which is LR.

C. Feature Selection

Results obtained from the PCA feature selection method are shown in the “Fig. 9”.There explained variance is plotted against the number of components.159 components were selected as the optimum number of components from the elbow point of the graph.

Results obtained from the Correlation Feature selection method are shown in the “Fig. 10”. The accuracy of the model is plotted against the threshold value. Maximum accuracy was obtained at a threshold value of 0.015. Therefore that feature set was selected for the model training which consisted of 150 components.

From the graph, it was clear that the feature set obtained from the PCA feature selection method gives a higher accuracy. Therefore that feature set was selected. Then the models were tested with the full feature test and the PCA feature set.

The results obtained for the Arousal model are shown in the “Fig. 12” Using the feature set obtained from the PCA method, we could get an improvement of 6.40 in R2 score and 2.37 in accuracy. Also, we could reduce the RMSE by 9.24 when compared to the model with the full feature set. Similarly, for the Valance model, we could get an improvement of 5.84 in R2 score and 4.94 in accuracy. Also, we could reduce the RMSE by 4.94 as shown in the “Fig. 13” Since the feature set obtained from the PCA methods gives higher performance in the model compared to the full feature set, that feature set

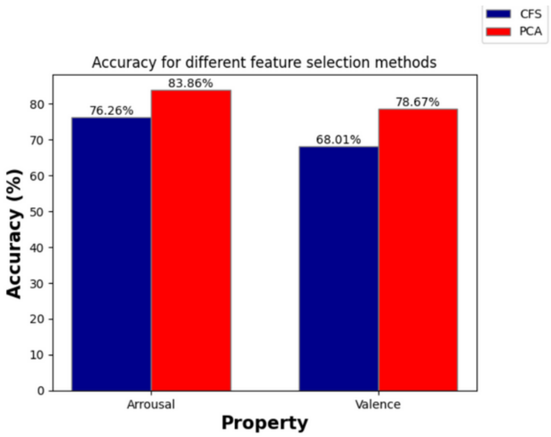

Fig. 11. Accuracy of the Models with Feature sets from PCA and CFS

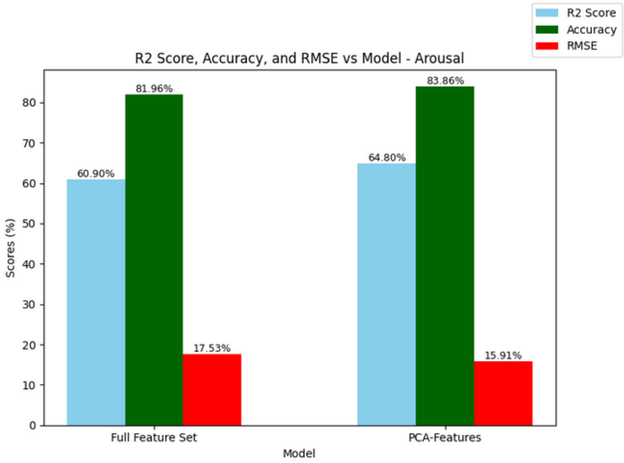

Fig. 12. Performance of Arousal model with full feature set and PCA feature set

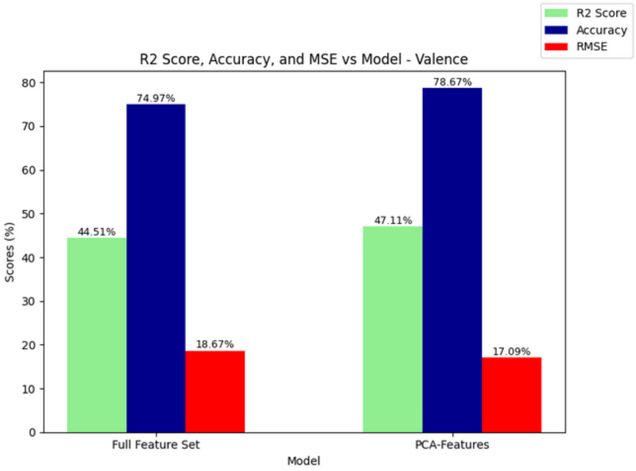

Fig. 13. Performance of Valance model with the full feature set and PCA feature set

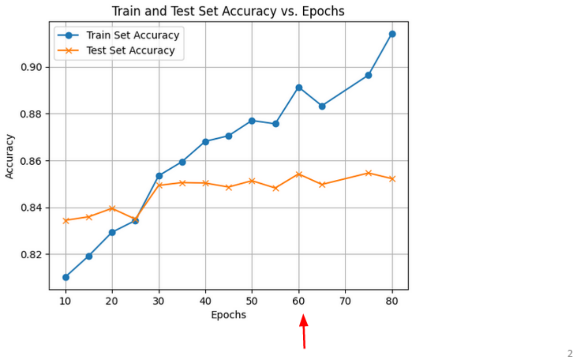

Fig. 14. Test and Train accuracies vs no.of epochs for Arousal model

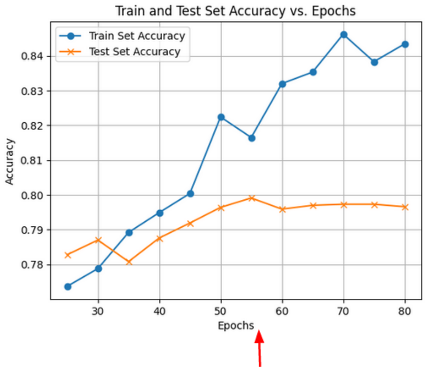

Fig. 15. Test and Train accuracies vs no.of epochs for Valence model

was selected for further improvements of the model.

D. Parameter Tuning

Under the parameter tuning optimum number of epochs, and learning rate were found. The accuracy of the model is plotted against the number of epochs for both the training and test datasets. From the graphs the optimum number of epochs of 60 for the arousal model and 55 for the valance model were selected as shown in the “Fig. 14” and “Fig. 15” respectively.

The optimum learning rate is obtained from the graph shown in the “Fig. 16” which plots the MSE vs learning rate of the validation dataset. The learning rate of 0.001 was selected since it gives the minimum MSE.

E. The Effect of Dense Layers

Several dense layers were added to the BILSTM model to improve the accuracy of the arousal model and valence model, the results obtained are shown in “Table. VI”. and “Table. VII”. respectively. After analyzing the results from those tables, the dense combination of 512,256,128,64,32 was taken for the finalized model.

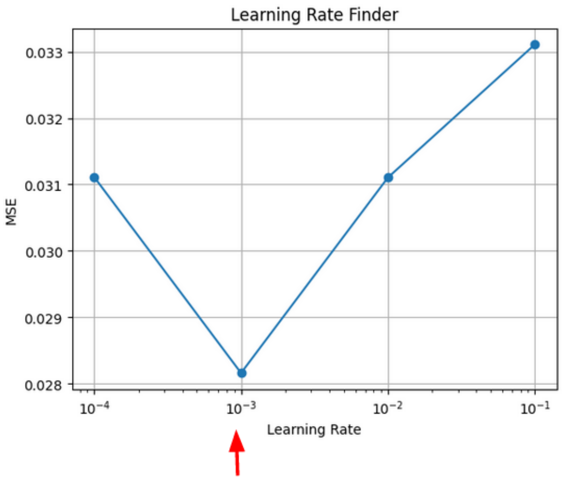

Fig. 16. learning rate vs MSE

TABLE VI
R2 SCORE, ACCURACY, AND RMSE VALUES OBTAINED FOR THE DIFFERENT DENSE LAYER COMBINATIONS - AROUSAL MODEL

Dense Layers	R2 Score (%)	Accuracy (%)	RMSE (%)
512	64.80	83.86	15.91
512,256	65.88	83.94	15.66
512,256,128	68.82	84.91	14.97
512,256,128,64	69.93	85.20	14.71
512,256,128,64,32	69.81	85.32	14.72
1024,1024,1024,1024,1024	66.17	84.15	15.59

TABLE VII
R2 SCORE, ACCURACY, AND RMSE VALUES OBTAINED FOR THE DIFFERENT DENSE LAYER COMBINATIONS - VALENCE MODEL

Dense Layers	R2 Score (%)	Accuracy (%)	RMSE (%)
512	47.11	78.67	17.09
512,256	46.45	78.35	17.19
512,256,128	48.28	78.91	16.90
512,256,128,64	51.22	79.51	16.41
512,256,128,64,32	51.79	79.82	16.31
1024,1024,1024,1024,1024	49.61	79.07	16.68

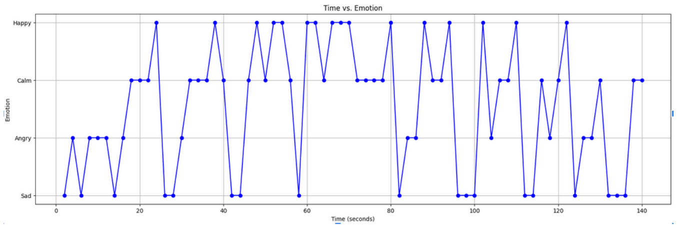

Fig. 17. Dynamic Emotional change of the song

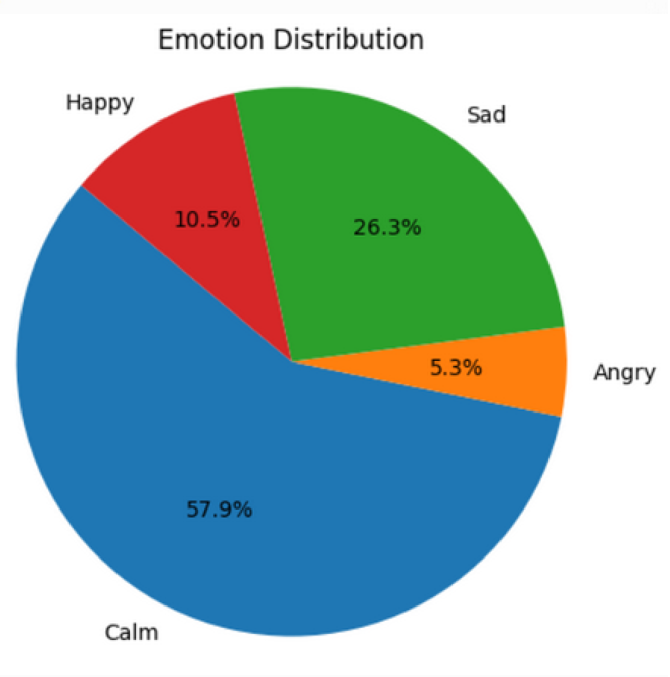

Fig. 18. Emotion Distribution

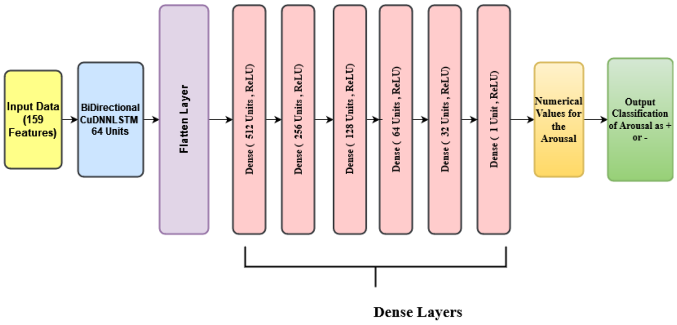
Input Data (159 Features)
BiDirectional CuDNNLSTM 64 Units
Flatten Layer
Dense (512 Units , ReLU)
Dense (256 Units , ReLU)
Dense (128 Units , ReLU)
Dense (64 Units , ReLU)
Dense (32 Units , ReLU)
Dense (1 Unit , ReLU)
Numerical Values for the Arousal
Output Classification of Arousal as + or -
Dense Layers

Fig. 19. Architecture for Arousal model

F. Model Architecture

After finalizing the parameters and configurations, the final models of arousal and valence are obtained as in the "Fig. 19" and "Fig. 20" respectively.

The final model comparisons for the arousal and valence models are shown in the figures 'Fig. 21" and "Fig. 22" respectively. So, at the end of our research, we were able to improve the accuracy of the arousal model by 10.39% and the R2 score by 25.52% compared to the baseline model. For the valence models improvements in accuracy and R2 score are 12.49% and 62.91% respectively. The equation that was used to obtain the percentage improvements is shown in "Fig. 23".

G. Implementation Outcomes

From the finalized models, we analyzed the dynamic emotional changes for a song."Fig. 17" shows how emotion changes in the song Shootout - by Izzamuzzic and Julien

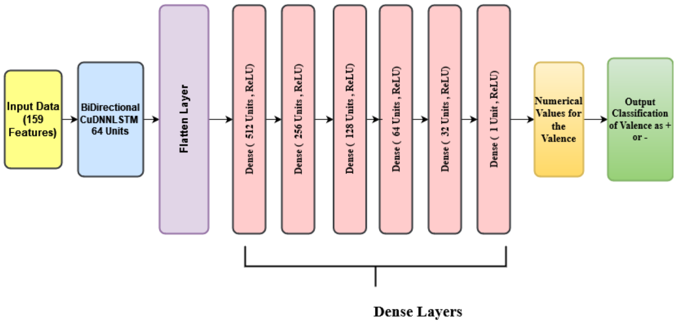
Input Data (159 Features)
BiDirectional CuDNNLSTM 64 Units
Flatten Layer
Dense (512 Units , ReLU)
Dense (256 Units , ReLU)
Dense (128 Units , ReLU)
Dense (64 Units , ReLU)
Dense (32 Units , ReLU)
Dense (1 Unit , ReLU)
Numerical Values for the Valence
Output Classification of Valence as + or -
Dense Layers

Fig. 20. Architecture for Valence model

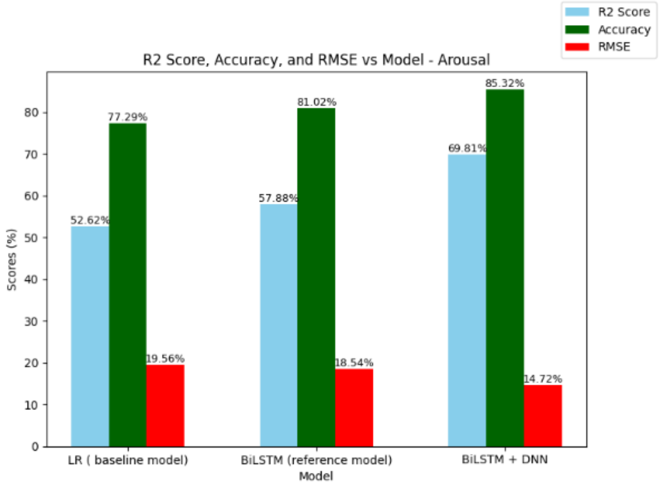

Fig. 21. Model Comparison - Arousal

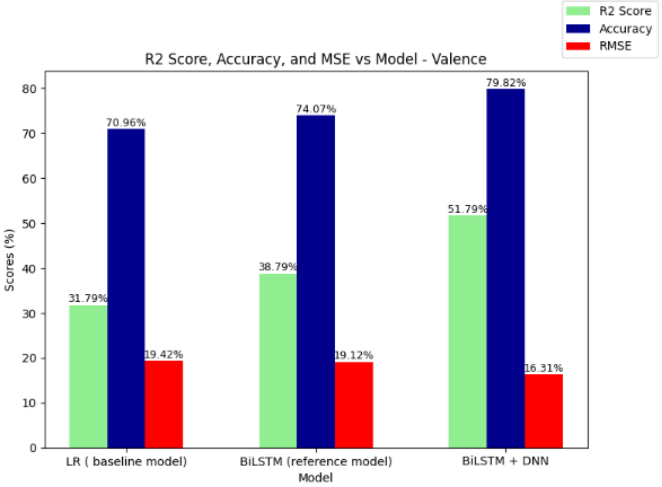

Fig. 22. Model Comparison - Valence

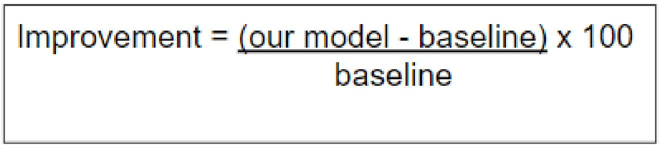
$$\text{Improvement} = \frac{(\text{our model - baseline})}{\text{baseline}} \times 100$$

Fig. 23. Equation to get the percentage improvement

Marchal. Overall emotional distribution is indicated as shown in the “Fig. 18”

VI. Conclusion

As a conclusion of our research, we found that the fact that systems that use deep neural network concepts have a higher accuracy compared to the systems that use only traditional machine learning algorithms. Also, it can analyze more complex music features to determine the emotional state more precisely. On the other hand, we found that Dynamic MER models are more accurate than the static MER models since they analyze the emotion of the music over the whole period. Also, the systems that use hybrid models rather than a single model display more accuracy. Therefore overall Dynamic Music Emotion Recognition Systems have a great impact on improved Music Recommendation Systems, Music Therapy, and enhanced music understanding since it recognizes changing emotions throughout a song. These are some major application areas of our research.

References

[1] Ja-Hwung Su, Tzung-Pei Hong1,Yao-Hong Hsieh, and Shu-Min Li, “Effective Music Emotion Recognition by Segment-based Progressive Learning”, October 2020.

[2] Kin Wai Cheuk, Yin-Jyun Luo, Balamurali B. T, Gemma Roig, and Dorien Herremans, “Regression-based Music Emotion Prediction using Triplet Neural Networks”, November 2020.

[3] Y.-H. Yang, Y.-C. Lin, and Y.-F. Su is with the Graduate Institute of Communication Engineering, National Taiwan University “A Regression Approach to Music Emotion Recognition”, September 20, 2007

[4] San Diego, La Jolla University of California, ”Comparison and Analysis of Deep Audio Embeddings for Music Emotion Recognition”,13 - April - 2021

[5] Xinyu Yang,Yizhuo Dong ,Juan Li “Review of data features-based music emotion recognition methods,”, November 2016

[6] Changfeng Chen, Qiang Li, “A Multimodal Music Emotion Classification Method Based on Multi-feature Combined Network Classifier”, August 2020

[7] Ye Ma, XinXing Li, Mingxing Xu, Jia Jia and, Lianhong Cai, “Multi-scale Context Based Attention for Dynamic Music Emotion Prediction”, October 2017.

[8] Haishu Xianyu, Xinxing Li, Wenxiao Chen, Fanhang Meng, Jiashen Tian, Mingxing Xu, and Lianhong Cai, “SVR Based Double-scale Regression for Dynamic Emotion Prediction in Music ”, 2016.

[9] Xinxing Li, Jiashen Tian, Mingxing Xu, Yishuang Ning, and Lianhong Cai,“DBLSTM - Based Multi-scale Fusion for Dynamic Emotion Prediction in Music ”.

[10] Xinxing Li, Haishu Xianyu, Jiashen Tian, Wenxiao Chen, Fanhang Meng, Mingxing Xu, and Lianhong Cai, “A Deep Bidirectional Long Short-Term Memory Based Multi-scale Approach for Music Dynamic Emotion Prediction ”, 2016.

[11] Yu Xia and Fumei Xu, “Study on Music Emotion Recognition Based on the Machine Learning Model Clustering Algorithm”, October 2022.

[12] Na He, Sam Ferguson from School of Computer Science, Faculty of Engineering and IT, University of Technology Sydney, Ultimo, NSW 2007, Australia.”Music emotion recognition based on segment-level two-stage learning”, 25 - April - 2022

[13] Xinxing Li, Jiashen Tian, Mingxing Xu, Yishuang Ning, and Lianhong Cai, “DBLSTM - Based Multi-scale Fusion for Dynamic Emotion Prediction in Music ”.6 June 2022

[14] Jacek Grekow from Faculty of Computer Science, Bialystok University of Technology, Poland. ”Music Emotion Recognition Using recurrent neural networks and pre-trained models”,08 August 2021

[15] S.Lalitha, D.Geyasruti, R.Narayanan, M.Shravani ,”Emotion Detection using MFCC and Cepstrum Features”, 20 - October - 2015.

[16] Xinxing Li, Jiashen Tian, Mingxing Xu, Yishuang Ning, and Lianhong Cai,“DBLSTM - Based Multi-scale Fusion for Dynamic Emotion Prediction in Music ”.

[17] Xinxing Li, Jiashen Tian, Mingxing Xu, Yishuang Ning, and Lianhong Cai,“DBLSTM - Based Multi-scale Fusion for Dynamic Emotion Prediction in Music ”.

[18] Ala Saleh Alluhaidan, Oumaima Saidani, Rashid Jahangir, Muhammad Asif Nauman, “Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network ”, April 2023

[19] Jing Yang,“A Novel Music Emotion Recognition Model Using Neural Network Technology” , September 2021

[20] Renato Panda, Ricardo Malheiro , Rui Pedro Paiva“Novel audio features for music emotion recognition”, 2018

[21] Nattapong THAMMASAN, Koichi MORIYAMA, Ken-ichi FUKUI, and Masayuki NUMAO “Continuous Music-Emotion Recognition Based on Electroencephalogram”, April 2016

[22] Serhat Hizlisoy, Serdar Yildirim, Zekeriya Tufekci “Music emotion recognition using convolutional long short term memory deep neural networks,”, 2020