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. So this literature review focuses on the approaches, strategies, and challenges involved in developing successful MER systems as it examines the state of the art in music emotion recognition research.**

***Index Terms*—Music Emotion Recognition, MER, emotional content analysis, music classification, feature extraction, machine learning algorithms**

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 which are included in the music compositions like melody, tempo, rhythm, harmony, timbre and 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 much useful in the media and entertainment sector since they can improve the audiovisual experience by analyzing the emotional content of the music.

So our aim through this literature review is to find the methods and technologies used in previous research works on MER, focusing on their advantages and drawbacks and potential areas for development. Then to improve the efficiency and the precision from the foundation laid by previous researchers.

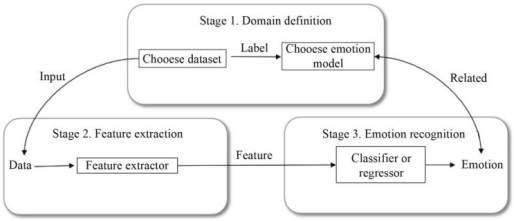

Fig. 1. Music Emotion Recognition framework

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 definition stage, then 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

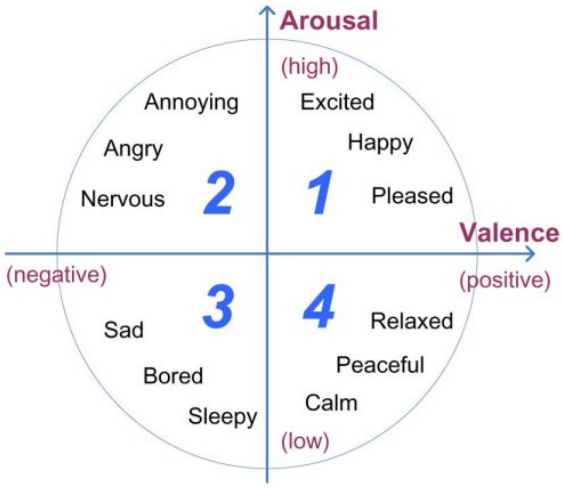

Fig. 2. Thayer's arousal-valence emotion plane

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.

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

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. III" and "Table. IV" show the summary of the literature review.

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 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

TABLE III
SUMMARY OF THE LITERATURE REVIEW

Reference	Dataset	Training Models	Evaluation	Result Measures	Results	Dynamic MER	Hybrid Models	Deep Learning
[1] 2020	CAL500	CNN, SVM	parameter settings, against other classification models	precision, recall, and F-measure,	results outperformed state-of-the-art methods	✗	✗	✓
[2] 2020	DEAM, MediaEval 2013	TNN, SVR, GBM	against baseline models with different dimensionality reduction methods	R2 score	Other dimensionality reduction techniques, such as principal component analysis (PCA) and autoencoders (AE), were surpassed by the TNN method.	✗	✓	✓
[3] 2008		MLR, SLR ,Ada Boost.RT	Performance of regression by the tenfold cross-validation technique, Evaluate the consistency of the ground truth in two ways Evaluate the prediction accuracy of different regression algorithms in terms of R2	R2 statistics	PC RRF directly affects arousal and valence , the resulting percentages are 58.3% for arousal and 28.1% for valence.	✗	✗	✗
[4] 2021	Youtube-8M ,4Q Audio Emotion	VGGish, SVM, NB, MLP, CNN, RNN	Performance Comparison with Baseline MER models, Against different data sets	F - Score R2 - Score	Results of Classification for Each Quadrant , The classification report of the bi-modal emotion dataset using the L3-Net embedding and SVM classifier.	✓	✗	✓
[6] 2020	Last. fm tag subset of the million song dataset	Multiple convolution kernels in CNN for 2D feature extraction, BiLSTM	Against other classification models for lyrics and audio	Accuracy of different lyrics classification models Accuracy of different multimodal fusion methods	Method outperformed single model classifications	✗	✓	✓
[7] 2017	MediaEval 2015	LSTM	against other regression models	Root Mean Square Error (RMSE)	method outperformed most of the models	✓	✗	✓
[8] 2016	MediaEval 2015	DS - SVR	Against baseline models	Mean Average Error (MAE) RMSE	method outperformed baseline models	✓	✗	✓

TABLE IV
SUMMARY OF THE LITERATURE REVIEW

Reference	Dataset	Training Models	Evaluation	Result Measures	Result	Dynamic MER	Hybrid Models	Deep Learning
[9]	MediaEval 2015	DBLSTM, triangle filter, MLR, SVR, ELM, ANN	different sequence lengths against the order of applying post-processing and fusion units	RMSE	the impact of sequence length on the DBLSTM model's performance during the training and prediction phases , The ideal result came from post-fusion processing.	✓	✓	✓
[10] 2016	MediaEval 2015	DBLSTM, ELM	against other regression models, different multi-scale fusion methods and different sequence lengths	RMSE	Multiscale fusion improved the performance, DBLSTM outperformed other models, and sequence length had an impact on performance.	✓	✓	✓
[11] 2022	MediaEval	SVR + RBRF (hybrid model)	against other regression and classification algorithms	MAE R2 score	Outperformed traditional methods	✗	✓	✗
[12] 2022	PMEmo All Music	CNN-based autoencoder model, BiLSTM mode	Evaluate the whole model by running tenfold cross-validation and obtaining the average performance	F1-score	The best valence results are shown by the PMEmo dataset, 1-s segment.	✓	✓	✓
[13] 2022	MIDI dataset	LSTM, CNN, HMM, MCMM, RL-RNN	against the algorithm's accuracy	Recall Rate	The algorithm in this paper has a better effect, with higher generalization, stability	✓	✗	✓
[19] 2021	AMG1608 CAL MIRER MediaEval-2015	Artificial bee colony (ABC) algorithm to improve the structure of BP neural network	Against other Algorithms	MAE RMSE R2	The BP model's results for emotion recognition outperform those of SVM, KNN, and GMM.	✗	✗	✗
[20] 2018	Created public dataset of 900 audio clips	SVM Baseline+novel Features	Novel Audio Features Were used	F1-Score Recall Precision	Resulting feature set (base-line+novel) outperforms others	✗	✗	✗
[22] 2021	A new Turkish emotional music database	LSTM, DNN	utilize a range of feature sets, By feeding several information sources into the CNN layer, features are obtained, MFCCs and log-mel filterbank energies	Accuracy precision recall F-measure	Performance increased with the proposed model compared to SVM, KNN, and RF after applying CFS	✗	✗	✓

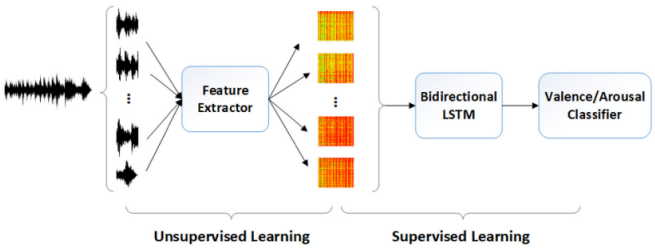

Fig. 3. : The two-stage learning approach uses a supervised learning model as an emotion recognizer and an unsupervised learning model as a segment-level feature extractor.

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 paper "Music Emotion Recognition based on Segment-Level Two-Stage Learning" [12], Na He and San Ferguson present an innovative two-part framework (Figure 3). 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.

Limitations of single network classification models in MER are presented in the paper "A Multimodal Music Emotion Classification Method Based on Multi-feature Combined Network Classifier" [6]. It suggests a new method that uses CNN-LSTM (convolutional neural networks-long short-term memory) models for audio and lyrics-based emotion classification. Here CNN component is used.BiLSTM component is used for the serialization process. The DNN layer in the model synthesizes feature information and enhances the fusion of emotional information. The authors have used some preprocessing techniques like fine audio segmentation, pure background sound extraction, and human voice separation to optimize the dataset. The overall accuracy of the audio classification has increased due to this combined classifier model and it also overcomes the limitations of the single-feature model.

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.

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 by them. 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 with combined CNN for feature extraction and LSTM + DNN for classification. This outperforms other classifiers like k-nearest neighbor (k-

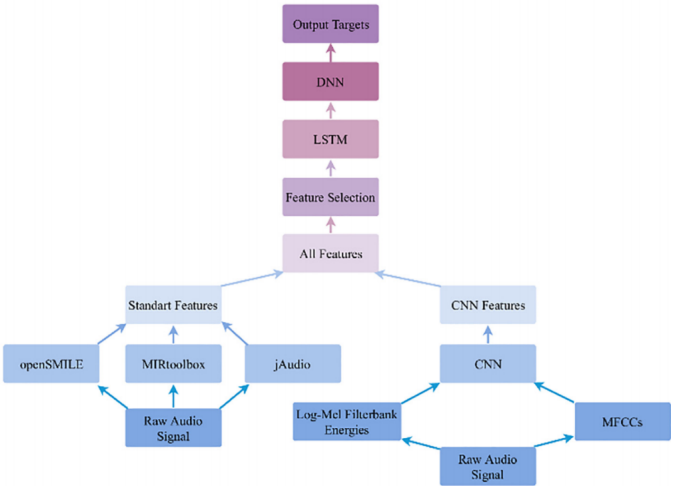

Fig. 4. The architectures of convolutional long short-term memory deep neural network.

NN), support vector machine (SVM), and Random Forest, achieving a higher overall accuracy through 10-fold cross-validation."Fig. 4" represents the overall procedure followed in this research paper.

In the paper," Effective Music Emotion Recognition by Segment-based Progressive Learning" by Yao-Hong Hsieh, Ja-Hwung Su, Tzung-Pei Hong, and Shu-Min Li [1], proposed a music emotion recognition algorithm, which uses Deep Learning (DL) and Support Vector Machine (SVM). The CAL500 dataset was used.

First, they decomposed music into a set of segments and transformed a music piece into a Mel-Frequency spectrum image. Then, a convolutional neural network (CNN) was used to model and recognize the images. After that, the outputs from the CNN were sent to the SVM model. Finally, the SVM model predicted the top k emotions for music. Using this method, they achieved a good improvement in contrast to the state-of-the-art methods in music emotion recognition.

The paper "Comparison and Analysis of Deep Audio Embeddings for Music Emotion Recognition," by San Diego and La Jolla at the University of California, [4] focused, on the effectiveness of two methods, L3-Net and VGGish.Here they have used four different labeled data sets to train the model. The Youtube-8M, 4Q Audio emotion data sets were used to test the model's performance. In this research, authors have compared the performance with some measurements such as accuracy level, recall, precision, and F1 score. By using those results they have measured the effectiveness of the VGGish model and L3-Net model. Their target is to identify the strengths and weaknesses of the L3-Net and VGGish deep audio embedding models. Here researchers have followed two major steps. In step 1, VGGish or L3-Net model is used to extract embeddings from a given song. These embeddings show the acoustic features of the song. Actually, those models are trained to capture and encode relevant information about the given music. In step 2, those extracted deep embeddings are used by a selected classification model to predict the emotion of the music. Here also they have used Thayer's arousal-valence emotion plane to identify the corresponding emotion according to AV values. This classification model has been trained by a labeled data set that consists of specific audio features.

The research paper, "Novel Audio Features for Music Emotion Recognition" [20] introduces various novel audio features for music emotion recognition. Melodic characteristics are one of those features. This includes MIDI note statistics, register distribution and ratios of pitch transitions. Note intensity statistics, distribution, and transitions are considered under dynamic features. Rhythmic features like note duration statistics and distribution, as well as musical texture features like musical layers statistics, distribution, and transitions, are also introduced in this work.SVM classifier is used for the testing with combined novel features and baseline features. Cross-validation techniques are also used here. The results show that using the novel features enhances the F1-score by 9% to reach 76.4% compared to using baseline-only features. Here they have created a public dataset of 900 audio clips annotated with different emotions. Russell's model is used with different emotion quadrants. This new dataset also provides genre, artist, and emotion tags for multi-label classification.

The paper "Study on Music Emotion Recognition Based on the Machine Learning Model Clustering Algorithm", by Yu Xia and Fumei Xu [11], developed a regression-based music emotion classification system. The dataset used in this study was in the MediaEval database. Three algorithms were used in the training part to obtain the regression model, and they are polynomial regression, support vector regression, and k-plane piecewise regression. In the testing part, the input music data was regressed and predicted to obtain its arousal and valence values and then classified. The system performance was considered by classification accuracy. According to their results, it can be seen that the combined method of support vector regression and k-plane piecewise regression has high accuracy, compared to using one algorithm alone.

Here, regression models of the arousal and valence values were obtained separately. In this study, support vector regression and RBFR were combined in the subsequent experiments. Support vector regression was used to obtain the arousal regression model RBFR was used to obtain the valence regression model and then observed the classification accuracy of the music emotion classification system, and it was observed that it had a certain improvement compared with the two methods alone. Also in this study, they have proved that the regression classification has a higher accuracy than the SVM classification. They applied different dimensionality reduction methods like PCA and Relief algorithms on features and tested the accuracy. It can be seen that the proposed hybrid classifier performed well with the Relief dimensionality reduction method.

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 as "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 "A Regression Approach to Music Emotion Recognition" [3], Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen, University of National Taiwan University focused to recognize music emotion. In this case, they have taken the Music Emotion Recognition as a regression problem. The final target was to detect the average arousal and valence values for a particular song. If we know the arousal and valence values we can use Thayer's arousal-valence emotion plane (shown in Figure 2) to identify the corresponding emotion according to AV values.

In this research, researchers have used three regression models. such as Support vector regression (SVR), multiple linear regression (MLR), and AdaBoost.RT (BoostR) to train the data.MLR is used as a baseline model due to the simplicity of this model. BoostR is a nonlinear regression algorithm. It is used for iterative processes, training multiple regression trees, and combining their prediction results. Finally, when we consider the research results, they have achieved, 78.4% of arousal accuracy and 21.9% of valence accuracy as average.

The literature review named "Review of Data Features-Based Music Emotion Recognition Methods" [5], points out the importance of using a combination of different data features during the modeling phase of MER. Some challenges of MER systems like selecting suitable standards for feature selection, and finding the regulating variations of emotions are discussed here. To overcome these challenges, it emphasizes 3 aspects of different data features such as utilizing music features only, ground truth data only, and their combination.

Also, this paper discusses some drawbacks of using only the music features since different music segments can convey different emotions though they have the same music features. Hence it can lead to confusion and decrease the accuracy. Furthermore, using solely ground truth data can result in unreliable results due to the randomness and uncertainty of human annotations. Due to the above reasons, researchers suggest using a combination of the above data types.

The research paper,"A Novel Music Emotion Recognition Model Using Neural Network Technology" [19] focuses on major challenges faced by existing(MER methods. Those are, due to the constantly changing emotional nature of the music, accurately expressing those based on the entire music is challenging, and analyzing the emotions solely based on pitch, length, and intensity of notes fails to capture the soul and connotation of the music. To overcome those challenges, they suggest an improved backpropagation (BP) algorithm neural network by integrating the artificial bee colony (ABC) algorithm. It optimizes the weights and thresholds of the BP neural network and enhances its global search ability Experimental results on public music datasets demonstrate that the proposed MER method outperforms other comparative models in terms of recognition effectiveness and speed.

Also, the research focuses on the classification of the acoustic features of music in a combined form to enable accurate emotion classification. Through experiments, it is determined that the classification of music features based on short-term energy, short-term average amplitude, short-term autocorrelation function, short-term zero-crossing rate, frequency spectrum, amplitude spectrum, and phase spectrum yields better results.

C. Conclusion

After analyzing the previous research works more deeply, we observed the fact that the systems which are using DNN, have a higher accuracy compared to the systems that use only the traditional machine learning algorithms. Also, it has the ability to analyze more complex music features to determine the emotional state of that more preciously. On the other hand, we found that Dynamic MER models are more accurate than the static MER models since it analyzes the emotion of the music over the whole time period. Also, the systems which have used hybrid models rather than a single model display more accuracy.

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