International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 13 Issue: 08 | Aug 2026 www.irjet.net p-ISSN: 2395-0072

AUTOMATIC MUSIC MOOD DETECTION USING A HYBRID CNN-LSTM MODEL

SHINDERPAL KAUR[1], Er. PARAMJIT KAUR[2]

[1]*Department of Computer Science and Engineering, Baba Farid College of Engineering and Technology, Bathinda, India*
[2]*Department of Computer Science and Engineering, Baba Farid College of Engineering and Technology, Bathinda, India*

Abstract - *People sometimes have the capacity to express their emotions through music which is unparalleled. Because of the explosion of digital music platforms, intelligent systems which are promising to automatically detect as well as categorize Emotional/Musical contents highly demanded. Music Emotion Recognition (MER) is the challenging and under defined task of recognizing emotions from the music. Here, we introduce a deep learning system for automatic mood recognition in music by combining CNN and LSTM network, which is a hybrid deep learning system. The proposed classification of music into four emotions, happy, sad, calm and energetic, is trained and evaluated on the DEAM dataset. The experimental results show the CNN-LSTM model suggested by the experiments can achieve 88% accuracy. This system is effective for recommendation systems, personalised playlist generation and music streaming applications.*

Key Words: Automatic Music Mood Detection, Music Emotion Recognition, Deep Learning, CNN-LSTM, DEAM Dataset, Audio Classification, Emotion Classification.

1. INTRODUCTION

One of the strongest mediums to express emotions in human life is music. As a result of the fast growth of music platforms like Spotify, Jio Saavn, Gaana, and YouTube Music, people have access to millions of songs each day. However, classifying and retrieving music according to its mood is both subjective and time-consuming process. Music may arouse a variety of feelings. Prior to listening to a song, music emotion recognition (MER) [1] may forecast the listener's emotional state. Music Emotion Recognition (MER) is an artificial intelligence (AI) subfield of Music Information Retrieval (MIR) that aims to recognize and classify the emotions expressed in musical works. Several machine learning algorithms are combined with a wealth of music information derived from digital audio samples to create the basis of music emotion recognition systems [2]. The most common approach to the MER problem is with machine learning techniques based on features which are derived from the acoustics of the sound, Parameters similar to the MFCC include chroma characteristics, spectral centroid and zero crossing rate. These approaches produce reasonable results; however, they do not account for the complicated nature of temporal and spectral dependencies between musical signals.

Deep learning techniques have recently been demonstrated to excel in audio categorisation problems**.** Mel-spectrogram data is temporal and sequential, fitting for convolutional neural networks (CNNs) to learn the "spatial" properties of it. However, LSTM networks are more suitable for modelling this data. The combination of CNN-LSTM hybrid was beneficial to detect musical moods based on the best performance of both models.

We propose an Automatic Music Mood Detection method, this is based on CNN-LSTM hybrid deep learning system. The four classes of mood used from the DEAM database (Happy, Sad, Calm, and Energetic) can be used in training and testing of the suggested model. In our experiments, we found that the proposed model shows 88% accuracy in predicting compared with the traditional machine learning techniques. The suggested system is best suited for online music streaming services.

This paper is organized as follows: The literature review is at Section 2. In Section 3, we go over the suggested methodology. The fourth section deals with results and analysis. The results of the study are summarized and suggestions for future research are presented in Section 5.

2. RESEARCH GAP

From the literature study, a substantial quantity of past research on Music Emotion Recognition becomes evident. Despite the efforts of Modran et al. to develop a trained neural network that will predict the effectiveness of music therapy, their work was based on therapist-aided selection Music Mood Online (3): Deep Learning Model based on CNN-LSTM (Convolutional Neural Network- Long Short-Term Memory) is used in the detection of Music Mood, while on the move.

Although the transformer-based HuBERT was shown to be effective in extracting the MFCC and STFT characteristics to classify the music mood, as described in [4], the use of hybrid CNN-LSTM models for automated music mood identification without relying on large scale pre-trained transformer models is still a fresh area of study.

Nevertheless, LSTM and CNN models of the Deep Learning offer better results.

Furthermore, most existing studies focus on either spatial feature extraction using CNN or temporal modeling using LSTM separately. Very few works have explored a hybrid CNN-LSTM architecture for music mood detection on the DEAM dataset. In spite of developing an explainable CNN architecture based on harmonic structure for automatic music emotion classification based on a theory-based approach, the authors mainly concentrated on model explainability and marketing aspects of the task. Hybrid CNN-LSTM approaches can help to achieve higher accuracy in automatic music mood classification but were not considered before [5]. Even though Qiao et al. [6] applied a CNN-SA-BiLSTM architecture for spatial and temporal feature extraction from EEG signals in order to perform music emotion recognition, the task was devoted to brain signal-based emotion recognition without consideration of hybrid CNN-LSTM models for audio-based automatic music mood classification. Even though Yu et al. [7] employed a CNN architecture together with IoT techniques in order to visualize Chinese music emotions by means of NetEase Cloud Music dataset, they did not consider hybrid CNN-LSTM models for audio-based automatic music mood classification. As opposed to Li et al. [8] who used an SVM method by machine learning for the recognition of musical emotions by timbre, rhythm, and pitch features extracted from the dataset of 499 songs, their study was related to classical machine learning and did not address the issue of hybrid CNN-LSTM networks in automatic feature learning and music mood recognition with the use of raw audio on diverse databases like DEAM. Computational Media Aesthetics and rule-based analysis of tempo and articulation were also used by Feng et al. [9] to suggest a mood-based music retrieval mechanism. The issue of detecting musical moods from raw audio stream using hybrid CNN-LSTM networks was not considered in their work, exploring only aspects of music theory.

That means that a model capable of achieving high accuracy wielding raw audio must be automatically learned at spatial and temporal features. To meet this requirement, we created a hybrid CNN-LSTM model to automatically classify songs into 4 mood categories.

3. METHODOLOGY

The proposed method for Automatic Music Mood Detection follows a systematic pipeline that collects the datasets, preprocesses the audio signal, extracts features, and classifies them in the hybrid CNN-LSTM model. It depicts the whole process of suggested system as illustrated in figure 1.

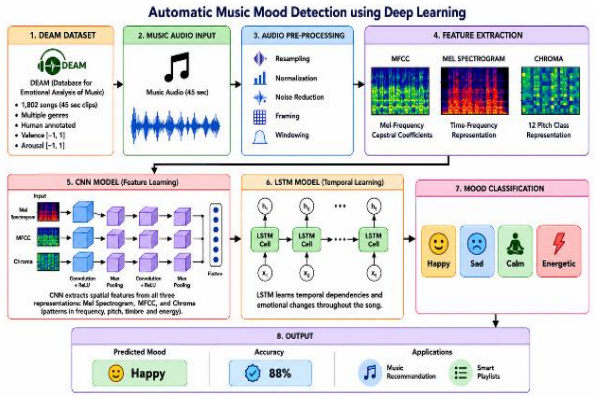

Fig. 1. Workflow of proposed system

3.1 Dataset Description

The dataset of DEAM is used in this investigation. DEAM consists of 1802 music pieces that are 45 seconds on average. All music pieces can be categorized as being in a happy, sad, calm, or energetic mood. While assessing the model's accuracy uses the remaining 20% of the dataset, 80% of the dataset is used for training.

3.2 Audio Preprocessing

The audio file is firstly made mono channel and resampled to 22050 Hz. Background noises and silences in audio are filtered out in order to enhance the quality of audio.

3.3 Feature Extraction

The preprocessed audio signal is then extracted into there are five sets of characteristics that are represented by The order of preference is MFCC, Tonnetz, Mel-Spectrogram and Spectral Contrast. The signal characteristics are referenced by MFCC (timbre), the tutorial is referenced by Chroma (PITCH CLASS), Mel-Spectrogram (time-frequency energy), Spectral Contrast (spectral peaks and valleys), and Tonnetz (TONALITY of the audio input). All the collected features are then concatenated and normalized into a 1D feature vector that acts as the input of 1D-CNN layer.

3.4 Proposed Hybrid CNN-LSTM Model

We propose architecture of sequential hybrid model in which 1D Convolutional layers are used to represent spatial information and Long Short-Term Memory (LSTM) layers are used to represent temporal information. Choi et al. [10] proposed a CNN based model for semantic segmentation with a Convolutional Recurrent Neural Network (CRNN) to embed CNN in their model. The presented use case did not have any functionality to classify a mood at this level; it only had music tagging and temporal feat attr summarized using the RNN and spatial features extracted. Tzirakis et al. [11] introduced a Convolutional Neural Network (CNN) based approach for feature extraction, supported by long short-

term memory (LSTMs) for temporal modelling and developed an end-to-end multimodal emotion recognition system which has been applied to the RECOLA dataset. Contrastingly, we propose a hybrid model of CNN-LSTM for speech-to-speech translation (S2S) with specific purpose for recognizing or transcribing speech for audio-based automatic music mood classification into four categories using the DEAM dataset.

The components of the model architecture are listed below:

1. The Conv1D layer is used to extract features from the audio stream by using Rectified Linear Unit (ReLU) activation function. In this case, a 3x3 size kernel is used when you have 64 filters.
2. To accelerate the training the second layer is batch normalisation layer which normalizes the output activations of the previous layer.
3. The third layer is MaxPooling1D layer to reduce the dimensionality of the data to the minimum by reducing with the max pooling with pool size.
4. The Dropout Layer: To prevent being overfitting, this layer will have a set percentage of neurones which will be randomly turned off during training. There is a 50% attrition rate.
5. LSTM Layer: There are 64 LSTM units in this layer.This aids the model in spotting trends in consecutive data over time.
6. Dropout Layer: Extra layer used to regularise the network and improve its generalisability with 0.5 dropout rate.
7. The seventh layer is called dense layer, and consists 32 ReLU nodes that are fully connected.
8. The Output Layer or Dense Layer is made up of four neurones with the activation function Softmax which are used for distinguishing/in recognizing four different states: energetic, joyful, sad, and tranquil.

This model of the neural networks for this collection of parameters has a total of 35748 parameters.

Adam optimiser with learning rate 0.0013.5 is used for training this model. After training, an evaluation takes place. We train the model for 30 epochs, giving our mini-batch size of 32. Patience value equals to 5, which indicates the usage of early stopping as a means to rescue appropriate weights from danger of overfitting and to properly follow weight loss. Some metrics like accuracy, precision, recall, and F1 Score are used to evaluate the performance of the model. Furthermore, a Confusion Matrix is also developed showing the classification of emotions.

4. RESULTS AND DISCUSSION

As a part of the assessment of this automatically generated mood categorisation model, this study has created an explainer network using the dataset DEAM and is used for deep learning. The DEAM dataset contains 1802 audio clips which are categorized into four types of moods, i.e., Energetic, Happy, Sad, and Calm.

The results attained from the DEAM dataset suggest that the proposed model has learned the temporal and spatial characteristics for mood classification–with success. Hershey et al. [12] found that CNN-based models perform well on big audio datasets, which is in line with our model's CNN component's outstanding performance. While Oramas et al. [13] demonstrated the benefits of a multimodal approach using audio, text, and user feedback, our proposed CNN-LSTM model achieves 88% accuracy using only audio spectrograms. This shows that for music mood classification, audio-only models can be effective when metadata is unavailable. The use of Constant-Q-Transform spectrograms for music emotion recognition has also been validated [14], where CNNs were used to classify basic emotions in movie soundtracks based on neuroscientific premises. In spite of its revolutionary role in revolutionizing the field of sequence modelling by Vaswani et al.'s attention-only Transformer, their work did not focus on investigating more attention-based models, rather they used Attention-based models in ART problem and other NLP tasks for automatic music mood classification from audio using the DEAM dataset [15].

4.1 Performance Analysis

The expert Hybrid CNN-LSTM Model provides the result with a correct explanation of 88% accuracy, but this will depend on the data applied in the model for the test. Table 1 shows F1 scores, accuracy and recall for each one of the courses.

Table 1: Classification Report of the Proposed Model

Class	Precision	Recall	F1-Score
Energetic	0.86	0.91	0.88
Happy	0.90	0.85	0.87
Sad	0.91	0.86	0.88
Calm	0.86	0.90	0.88
Accuracy			0.88

For both of these parameters, the average F1-Score value, indicated in Table 1 is expected to be high, meaning that the model works well in general. This is because Mel-spectrograms were able to have their spatial and temporal properties extracted using the CNN-LSTM approach.

It's obvious from Figure's that each of the four moods was fairly well represented by the model. The values along the diagonal prove that the classification was successful for all 4 types of moods.

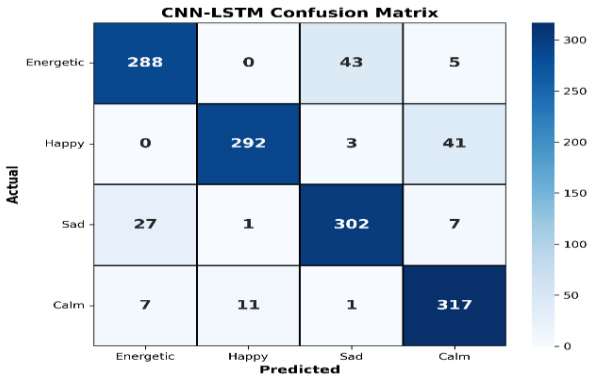

Fig. 2. Confusion Matrix of the Proposed Model

It is noted that the accuracy performance of the training and validation are displayed throughout the 30-epoch training period as shown in Figure 3. Both the graphs increase steadily and come to be about at 88% at the 30th epoch. The validation accuracy curve nearly matches the training accuracy curve, indicating good model learning, with no over fitting.

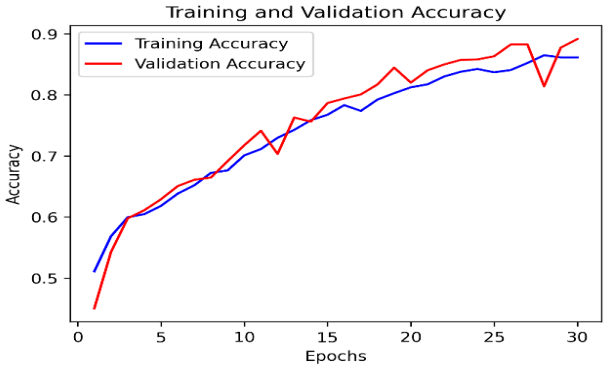

Fig. 3: Training and Validation Accuracy of the Proposed Model

The graph of the training loss and validation loss can be seen in Figure 4. After 20 runs, the losses are totaled from the training run and the validation run. Model performance shows that this model is able to perform well in generalisation as the training loss is not significantly different from the validation loss. The similarity in both the curves shows that there is no overfitting in the model because of Dropout and Early Stopping.

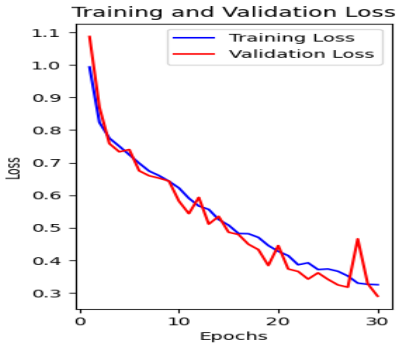

Fig. 4: Training and Validation Loss of the Proposed Model

The results show the effectiveness of extracting features via 1D CNN followed by classification through the temporal dependency extracted from LSTM applied in an autonomous music mood identification system.

5. CONCLUSION

The current study provides an efficient approach to Hybrid CNN-LSTM in automatic music mood detection through audio signals. The MFCC data are fed into a hybrid system that uses LSTM for learning about temporal dependencies and 1D CNN for extracting spatial information.The study demonstrates experimental results with an 88% classification accuracy rate for the The suggested model has the potential to attain four different mood types: busy, joyful, sorrowful, and serene. Under the average precision, recall and f1 score, the model is efficient across different scenarios with an average score of 0.88. As explained, Dropout and Early Stopping approaches do not cause the model to overfit, as it can be seen from the training and validation curves. The findings also show that the CNN and LSTM model is superior to the conventional ML algorithms when it comes to music emotion categorization. This is because the audio information can be used to train the model on local pattern and sequence information. Actually, there are plenty of applications in the real world, such as recommending products, making playlists of user's mood, and music analysis when therapeutic purposes. Future research could involve the application of this system to larger databases, more moods, and streaming audio signals. Data augmentation and attention methods could also be utilized in the future.

REFERENCES

1. Jandaghian, M., Setayeshi, S., Razzazi, F., & Sharifi, A. (2023). Music emotion recognition based on a modified brain emotional learning model. Multimedia Tools and Applications, 82(17), 26037-26061.
2. Lucia-Mulas, M. J., Revuelta-Sanz, P., Ruiz-Mezcua, B., & Gonzalez-Carrasco, I. (2023). Automatic music emotion classification model for movie soundtrack

subtitling based on neuroscientific premises: MJ Lucia-Mulas et al. Applied Intelligence, 53(22), 27096-27109.

3. Modran, H. A., Chamunorwa, T., Ursuțiu, D., Samoilă, C., & Hedeşiu, H. (2023). Using deep learning to recognize therapeutic effects of music based on emotions. Sensors, 23(2), 986.
4. Sun, Y. (2025). Feature centric based deep learning approach for music mood recognition with HuBERT transformer model. Scientific Reports.
5. Fong, H., Kumar, V., & Sudhir, K. (2025). A theory-based explainable deep learning architecture for music emotion. Marketing Science, 44(1), 196-219.
6. Qiao, Y., Mu, J., Xie, J., Hu, B., & Liu, G. (2024). Music emotion recognition based on temporal convolutional attention network using EEG. Frontiers in human neuroscience, 18, 1324897.
7. Yu, X. (2025). The impact of music visualization model by using internet of things techniques and deep neural network. Scientific Reports, 15(1), 39659.
8. Li, T., & Ogihara, M. (2003). Detecting emotion in music.
9. Feng, Y., Zhuang, Y., & Pan, Y. (2003, July). Popular music retrieval by detecting mood. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 375-376).
10. Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017, March). Convolutional recurrent neural networks for music classification. In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 2392-2396). IEEE.
11. Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of selected topics in signal processing, 11(8), 1301-1309.
12. Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., ... & Wilson, K. (2017, March). CNN architectures for large-scale audio classification. In 2017 ieee international conference on acoustics, speech and signal processing (icassp) (pp. 131-135). IEEE
13. Oramas, S., Nieto, O., Sordo, M., & Serra, X. (2017, August). A deep multimodal approach for cold-start music recommendation. In Proceedings of the 2nd workshop on deep learning for recommender systems (pp. 32-37).
14. Lucia-Mulas, M. J., Revuelta-Sanz, P., Ruiz-Mezcua, B., & Gonzalez-Carrasco, I. (2023). Automatic music emotion classification model for movie soundtrack subtitling based on neuroscientific premises: MJ Lucia-Mulas et al. Applied Intelligence, 53(22), 27096-27109.
15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.