JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Deep Learning-Based Emotion Recognition from Instrumental Hindi Music Signals

[1]Akanksha Gupta, [2] Anamika Singh,

[1]Research Scholar, [2]Assistant Professor
[1]Electronics and communication,
[1]LNCT University, Bhopal, India

Abstract : Music Emotion Recognition (MER) is an important research area within Music Information Retrieval that focuses on automatically identifying emotional characteristics conveyed by music. Although deep learning approaches have shown promising performance, most existing studies rely heavily on vocal or lyrical information and are largely based on Western music datasets. As a result, emotion recognition from instrumental Hindi music remains relatively underexplored. This paper presents a Convolutional Neural Network (CNN)-based framework for emotion recognition using instrumental (voice-removed) Hindi music from the MER500 dataset. Audio signals are converted into spectrogram representations using the Short-Time Fourier Transform and used as input to a CNN for automatic learning of discriminative spectral–temporal features. Experiments are conducted under two configurations: a five-category classification setting (*Devotional*, *Happy*, *Party*, *Romantic*, and *Sad*) and a four-category setting obtained by removing the *Happy* emotion class, which exhibits high ambiguity in instrumental music. The five-category experiment achieves an overall classification accuracy of 64% with a macro-averaged F1-score of 0.64, while excluding the *Happy* class improves performance to a test accuracy of 67.80% and a macro F1-score of 0.68. The results demonstrate that instrumental Hindi music contains meaningful emotional cues and emphasize the importance of emotion taxonomy design when performing emotion recognition without vocal or lyrical information. This work contributes to culturally diverse MER research and provides a foundation for future multimodal emotion recognition studies.

IndexTerms **- Music Emotion Recognition, Convolutional Neural Network, Instrumental Hindi Music, Deep Learning**

I. INTRODUCTION

Music has long been recognized as a powerful medium for expressing, conveying, and evoking human emotions. In everyday life, people often select and consume music based on their emotional state, such as happiness, sadness, relaxation, or excitement. With the exponential growth of digital music libraries and online streaming platforms, managing and organizing music content using traditional metadata such as artist, album, or genre has become increasingly insufficient. As a result, emotion-based music organization and retrieval has emerged as an important research direction, leading to significant interest in Music Emotion Recognition (MER) [1], [2].Music Emotion Recognition aims to automatically identify the emotional characteristics conveyed by a musical piece using computational techniques. MER is considered a subtask of Music Information Retrieval (MIR) and is inherently interdisciplinary, involving audio signal processing, machine learning, music psychology, and cognitive science. However, modeling emotions from music is particularly challenging due to the subjective nature of emotion perception, cultural influences, and the dynamic structure of musical signals [1], [3]. Unlike tasks such as genre classification, emotional labels often lack a single “correct” answer, making MER a complex and open research problem.

Early MER research primarily focused on traditional machine learning approaches, where emotional information was modeled using handcrafted audio features. These features typically included timbral descriptors (e.g., MFCCs), rhythmic patterns, harmonic features, and spectral statistics. Classification models such as Support Vector Machines, k-Nearest Neighbors, Gaussian Mixture Models, Random Forests, and shallow Artificial Neural Networks were widely used to map these features to emotion labels [1], [2]. Although these approaches demonstrated promising results, their performance strongly depended on the quality of feature engineering and domain expertise. Furthermore, handcrafted features often fail to capture the hierarchical and nonlinear relationships between musical structure and emotional perception [3], [4]. With the rapid development of deep learning (DL), MER research has gradually shifted toward data-driven representation learning. Deep learning models are capable of automatically learning high-level and discriminative features directly from raw or transformed audio signals, thereby reducing reliance on manual feature design. Among various deep learning architectures, Convolutional Neural Networks (CNNs) have gained particular attention in MER due to their effectiveness in modeling local spectral and temporal patterns from time–frequency representations

such as spectrograms and Mel-spectrograms. CNN-based approaches have consistently shown improved performance over traditional machine learning models in multiple MER benchmarks [5], [6].

Hindi music, in particular, exhibits rich melodic structures, diverse rhythmic patterns, and strong emotional expression influenced by lyrics, cultural context, and musical modes. These characteristics make Hindi music emotion recognition both challenging and distinct from Western music analysis. Despite its popularity and cultural significance, systematic deep learning-based MER studies focusing on Hindi music are still scarce. Moreover, balanced datasets that allow fair evaluation across multiple emotion categories are rarely explored. To address these limitations, this paper proposes a CNN-based deep learning framework for Music Emotion Recognition using the MER500 dataset [32], which consists of 500 Hindi songs evenly distributed across five emotion categories. By employing a deep learning approach, the proposed system automatically learns emotional representations from audio-derived inputs, overcoming the limitations of handcrafted feature-based models. The use of a balanced, language-specific dataset further enables a more reliable and meaningful evaluation of emotion recognition performance. The primary objective of this work is to investigate the effectiveness of CNN-based deep learning models for Hindi music emotion recognition and to provide insights into their advantages over traditional machine learning approaches. Through this study, we aim to contribute toward the development of robust MER systems that are better suited for language-specific and culturally diverse music collections.

II. PAST WORK

Music Emotion Recognition (MER), as a core subtask of Music Information Retrieval (MIR), was initially dominated by classical machine learning approaches. Early MER systems relied heavily on handcrafted feature engineering, where researchers manually extracted descriptors related to melody, harmony, rhythm, dynamics, pitch, and timbre from audio signals. These features were subsequently classified using traditional algorithms such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Random Forests (RF), and Gaussian Mixture Models (GMM) [1], [2]. Although these approaches demonstrated reasonable performance, their effectiveness was fundamentally constrained by the quality of handcrafted features. Several studies reported that classification accuracy often plateaued around 65–70%, indicating a "glass ceiling" effect [3]. This limitation arises due to the persistent semantic gap between low-level acoustic descriptors and high-level human emotional perception. As a result, traditional machine learning methods struggled to capture the complex, nonlinear, and hierarchical nature of emotional expression in music. To overcome these limitations, MER research has increasingly shifted toward Deep Learning (DL) paradigms, which unify feature extraction and classification into an end-to-end learning framework. Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), automatically learn hierarchical representations directly from raw audio waveforms or time–frequency representations like Mel-spectrograms. CNNs have proven particularly effective in capturing local spectral patterns, while RNNs and Long Short-Term Memory (LSTM) networks are capable of modeling long-term temporal dependencies and emotional evolution over time [4], [5].

Recent MER research has moved beyond conventional CNN and RNN architectures toward more advanced and integrated frameworks. Transformer-based models, incorporating self-attention mechanisms, have been introduced to focus on emotionally salient segments of music tracks and to model global temporal dependencies more effectively than traditional recurrent networks [6].Another major advancement is the emergence of large-scale pre-trained music encoders, such as the Music undERstanding model (MERT), which leverage self-supervised learning on massive music corpora. These models extract rich and transferable representations of timbre, harmony, and musical semantics, significantly outperforming traditional CNN-based systems when fine-tuned for MER tasks [7].Furthermore, modern MER systems increasingly adopt multitask learning strategies to jointly model categorical and dimensional emotion representations. This allows networks to learn shared emotional representations across heterogeneous datasets. In parallel, multimodal fusion approaches combine audio features with complementary modalities such as lyrics, music videos, and even physiological signals (e.g., EEG), resulting in improved robustness and emotion recognition accuracy [8], [9]. Recent studies have also explored specialized loss functions to enhance feature discrimination in deep models. Margin-based losses such as AM-Softmax, AAM-Softmax, and more recently AMCM-Softmax (Angular and Cosine Margin Combined Softmax) have been shown to improve intra-class compactness and inter-class separability, particularly for datasets containing emotionally similar music categories [10].

Despite the rapid progress in MER, most benchmark datasets—such as GTZAN, the Million Song Dataset (MSD), and MagnaTagATune—are heavily biased toward Western pop and classical music [2], [11]. Models trained on these datasets are optimized for Western musical constructs, including equal-tempered scales, major–minor tonality, and Western instrumentation. Consequently, their generalization to non-Western musical traditions remains limited. Focusing on Hindi and Indian music addresses this critical gap by introducing cultural and linguistic diversity into MER research. Indian music is fundamentally different from Western music in terms of structure, emotion representation, and performance practices. A defining characteristic of Indian Classical and Hindi music is the use of ragas, which are melodic frameworks designed to evoke specific emotional states, known as rasas. Unlike Western scales that often map emotions into binary categories (e.g., major–happy, minor–sad), a single raga can simultaneously convey multiple emotional dimensions such as devotion, awareness, tranquility, and intensity [12], [13]. This intrinsic complexity makes emotion recognition in Indian music a significantly more challenging task. In addition, Indian and Hindi music employs culture-specific instrumentation, such as sitar, tabla, tanpura, and harmonium, which produce distinct timbral and spectral characteristics not captured by models trained exclusively on Western instruments like guitar or piano [14]. These differences further emphasize the need for region-specific datasets and tailored deep learning architectures.

Recent studies have increasingly explored MER in Indian music domains using both traditional and deep learning approaches. Early work employed handcrafted rhythmic, timbral, and intensity-based features combined with clustering or classification techniques such as Fuzzy C-means and SVMs for Hindi music emotion categorization [15]. While informative, these methods suffered from limited scalability and generalization. More recent research has adopted deep learning architectures for Indian music

analysis. CNN-based models, including VGG16, ResNet18, and SqueezeNet, have been applied to spectrogram representations of Indian Classical Music (ICM), yielding encouraging results in emotion and raga classification tasks [16]. For Hindi popular music, CNN-based MER systems have demonstrated superior performance compared to traditional machine learning classifiers when classifying songs into multiple emotion categories based on Russell's valence–arousal model [17]. Specialized datasets such as JUMusEmoDB, consisting of sitar excerpts annotated for emotional content, have been introduced to facilitate Indian music emotion research [18]. Additionally, recent work on the MER500 dataset, which includes culturally specific labels such as Devotional, has explored advanced loss functions like AMCM-Softmax to enhance discrimination between emotionally similar classes in regional music [10]. Beyond emotion recognition, deep learning has also been applied to related Indian music tasks such as raga identification, automatic music transcription, and sequence modeling using LSTM networks. These developments contribute to the broader field of Traditional Music Deep Learning (TMDL), which aims to preserve cultural heritage while enabling intelligent music analysis systems [19].

III. RESEARCH METHODOLOGY

This section describes the complete research methodology adopted in this work for instrumental Hindi Music Emotion Recognition (MER). The proposed methodology is designed to closely align with the implemented Python-based experimental pipeline and consists of multiple sequential stages, including vocal removal, audio preprocessing, spectrogram generation, dataset construction, CNN-based model design, training strategy, and experimental evaluation under different classification settings.

3.1 Instrumental Data Preparation and Spectrogram Generation Using STFT

In the proposed study, all music tracks are preprocessed to remove vocal components before any feature extraction or model training. The motivation behind vocal removal is to eliminate the influence of lyrics and singing voice, thereby enabling the analysis of pure musical emotion conveyed through instrumental elements such as melody, harmony, rhythm, and timbre. All subsequent processing and experiments are performed exclusively on these instrumental audio signals[20] [21]. Each instrumental audio file is loaded using the *torchaudio* library. Since audio recordings may contain multiple channels, stereo signals are converted into a single-channel (mono) signal by averaging across channels. This step ensures channel consistency across the dataset and prevents channel-specific bias during training.

The time-domain audio signals are converted into a suitable representation for deep learning, Short-Time Fourier Transform (STFT) [22] is applied. Spectrograms are generated using an FFT window size of 1024 samples and a hop length of 512 samples, producing a power spectrogram with a power factor of 2.0. This configuration provides an effective balance between time and frequency resolution for music emotion analysis.The amplitude values of the spectrogram are converted into the logarithmic decibel (dB) scale using amplitude-to-decibel transformation. This enhances low-energy components and compresses the dynamic range, making subtle emotional cues more distinguishable. Each spectrogram is saved as a PNG image using a perceptually meaningful color map. Axis labels and padding are removed so that only spectral–temporal information is preserved. These spectrogram images form the primary input to the CNN model [23].

3.2 Dataset Construction and Label

The processed instrumental dataset is organized in a category-wise directory structure, where each subfolder corresponds to an emotion class. Emotion labels are automatically assigned numerical indices using an index-based encoding scheme. For each instrumental audio sample, a corresponding spectrogram image is generated and stored in a mirrored directory structure. The dataset is constructed as labelled pairs of (instrumental spectrogram image, emotion label), ensuring consistent and unbiased labelling throughout the dataset.

3.3 Data Splitting and Input Transformation

To ensure reliable performance evaluation, the dataset is divided into training (70%), validation (15%), and testing (15%) subsets using a stratified sampling strategy to preserve class balance across all splits. Before training, spectrogram images are converted to grayscale, resized to 128 × 128 pixels, and normalized to zero-centered values. These preprocessing steps standardize the input representation and improve training stability. In addition to the fixed split evaluation, a three-fold cross-validation strategy is employed in later experiments to assess the robustness and stability of the proposed model.

3.4 Convolutional Neural Network (CNN) Model Architecture

In this study, a Convolutional Neural Network (CNN) [24] [25] is employed to perform emotion classification from instrumental spectrogram images. CNNs are particularly suitable for music emotion recognition tasks due to their ability to learn hierarchical spatial patterns from time–frequency representations, such as spectrograms, which encode both spectral and temporal information.

The input to the CNN consists of grayscale spectrogram images of size 128 × 128 pixels, generated from instrumental Hindi music signals. The proposed CNN architecture comprises two convolutional layers followed by pooling operations and fully connected layers. The first convolutional layer applies 32 filters of size 3 × 3 with a stride of 1 and padding of 1, allowing the network to preserve spatial resolution while learning low-level features such as energy distribution, timbral texture, and local spectral variations. A Rectified Linear Unit (ReLU) activation function is used to introduce non-linearity, followed by a 2 × 2 max-pooling layer to reduce spatial dimensionality and improve translation invariance. The second convolutional layer consists of 64 filters of size 3 × 3, again followed by ReLU activation and max-pooling. This layer captures higher-level spectral–temporal patterns related to rhythm, harmonic structure, and instrumental characteristics that are strongly associated with musical emotions.

After the convolutional and pooling operations, the resulting feature maps are flattened and passed through a fully connected layer with 128 neurons, enabling the model to learn complex nonlinear relationships between extracted features and emotion classes. The final output layer consists of neurons equal to the number of emotion categories and produces raw class scores for emotion prediction. During training, these scores are optimized using a Softmax-based categorical cross-entropy loss [27]. Overall, the proposed CNN architecture provides an effective balance between model complexity and computational efficiency, making it suitable for emotion recognition from instrumental music spectrograms while reducing the risk of overfitting on a relatively small dataset.

IV. EXPERIMENTAL SETUP

This section describes the experimental configuration used to evaluate the proposed instrumental music emotion recognition framework, including dataset preparation, experimental settings, and implementation details.We systematically analyze emotion ambiguity and class separability in instrumental music emotion recognition, multiple experimental configurations are designed. First, a five-category classification experiment is performed using all available emotion classes to establish baseline performance. Subsequently, a four-category classification experiment is conducted by excluding the Happy emotion class, which was observed to exhibit significant overlap with other emotion categories in instrumental settings. Finally, a three-fold cross-validation experiment is carried out on the four-category dataset to evaluate the robustness and generalization capability of the proposed CNN-based framework while reducing dependency on a single train–test split.

For single-split experiments, the dataset is divided into training, validation, and testing subsets using a stratified sampling strategy to preserve class balance across all splits. Specifically, 70% of the data is allocated for training, 15% for validation, and the remaining 15% for testing. This partitioning strategy enables effective model learning, hyperparameter tuning, and unbiased performance evaluation. In addition to the fixed split evaluation, a cross-validation strategy is employed to further assess model stability. For three-fold cross-validation, the dataset is partitioned into three equal folds, where in each iteration two folds are used for training and the remaining fold is used for testing. This process is repeated until each fold has served as the test set exactly once, providing a more reliable estimation of model performance on limited data.

All experiments are implemented using the PyTorch deep learning framework [28][29]. Model training is performed for 30 epochs with a batch size of 16. The Adam optimizer [26] is employed with a learning rate of 0.001 to ensure efficient and stable convergence during training. After each epoch, the model is evaluated on the validation dataset to monitor generalization performance and detect overfitting. Model training and evaluation are carried out on available CPU hardware. The same experimental settings and hyperparameters are maintained across all experimental configurations to ensure fairness and reproducibility of the reported results.

The performance of the proposed CNN-based music emotion recognition system is evaluated using standard multi-class classification metrics to provide a comprehensive assessment of model behavior. Overall classification accuracy [30][31] is used to measure the proportion of correctly classified music samples across all emotion categories from equation (1). Precision is employed to evaluate the reliability of the predicted emotion labels by measuring the proportion of correctly predicted samples for each class using equation (2), while recall assesses the model's ability to correctly identify samples belonging to a specific emotion category from equation (3). The F1-score, defined as the harmonic mean of precision and recall, is used as a balanced performance measure, particularly suitable for datasets with class overlap and varying classification difficulty from equation (4). In addition to these metrics, confusion matrices are analyzed to examine inter-class misclassification patterns, and per-class accuracy is computed to evaluate class-wise recognition performance. Together, these evaluation metrics provide both global and fine-grained insights into the effectiveness and limitations of the proposed instrumental music emotion recognition framework.

$$\text{Average Accuracy} = \frac{\sum_{i=1}^{N} \frac{TP_i + TN_i}{TP_i + TN_i + FT_i + FN_i}}{100} \text{x} 100 \quad (1)$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \quad (2)$$

$$R_i = \frac{TP_i}{TP_i + FN_i} \quad (3)$$

$$F_i = \frac{2P_i R_i}{P_i + R_i} \quad (4)$$

V. RESULTS AND DISCUSSION

In the first experimental setting, the proposed CNN-based framework is evaluated on the complete five-category instrumental Hindi music dataset consisting of *Devotional*, *Happy*, *Party*, *Romantic*, and *Sad* emotions. This experiment establishes a baseline to assess whether emotional cues can be reliably extracted from **instrumental music alone**, without the influence of vocals or lyrics.

Table 5.1 Performance metrics for five-category instrumental MER

Emotion	Precision	Recall	F1-score	Accuracy (%)
Devotional	0.80	0.80	0.80	80.00
Happy	0.47	0.47	0.47	46.67
Party	0.60	0.60	0.60	60.00
Romantic	0.62	0.67	0.65	66.67
Sad	0.69	0.64	0.67	64.29
Overall	—	—	**0.64**	**64.00**

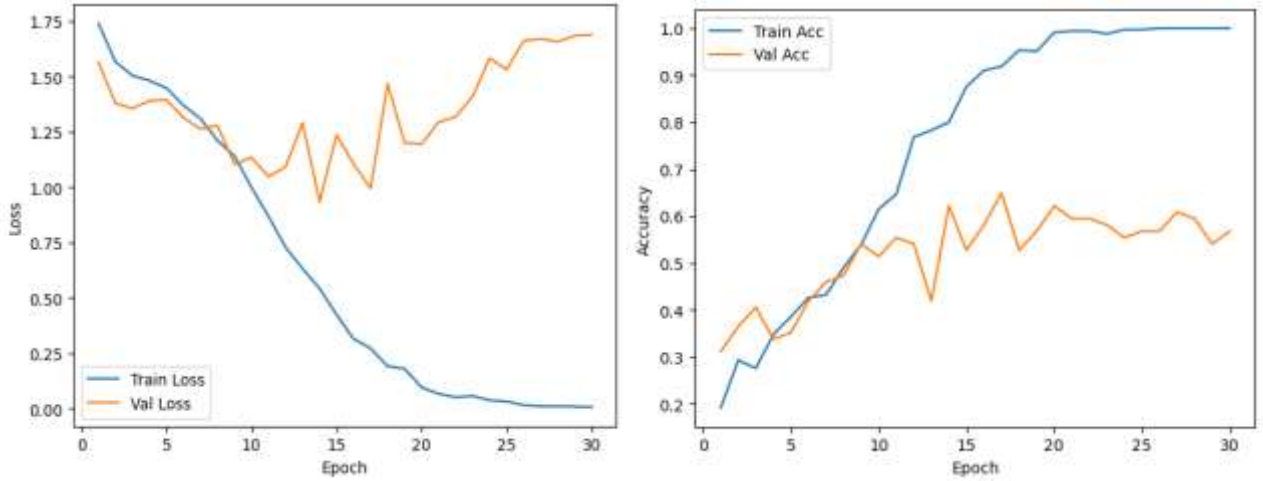

Figure 5.1 (a) Training and validation loss curves (b) Training and validation accuracy curves for the 5-category CNN-based instrumental MER model.

The training and validation performance curves indicate that the CNN effectively learns discriminative spectral–temporal representations from spectrogram images. Training accuracy consistently increases and approaches saturation, while validation accuracy stabilizes around **55–65%**, resulting in a noticeable generalization gap. From a theoretical perspective, this behaviour can be attributed to two primary factors: the limited size of the dataset and the high intra-class variability of musical emotions. Instrumental music often conveys emotions through subtle variations in tempo, harmony, and timbre, which are more difficult to model than explicit lyrical semantics.

On the unseen test set, the model achieves an overall classification accuracy of **64%** with a macro-averaged **F1-score of 0.64**. Class-wise analysis shows that *Devotional* emotion achieves the highest recognition accuracy (**80%**). This can be theoretically explained by the relatively stable musical structure of devotional music, which typically exhibits slower tempo, repetitive melodic phrases, sustained tonal centers, and limited dynamic variation. These characteristics produce consistent spectral patterns that are easier for CNN filters to capture.

In contrast, the *Happy* emotion exhibits the lowest recognition performance (**46.67% accuracy**). From a music cognition perspective, happiness in instrumental music does not correspond to a single acoustic pattern; instead, it overlaps strongly with *Party* (high energy, rhythmic intensity) and *Romantic* (major tonalities, melodic smoothness). Without vocal or lyrical cues, the CNN struggles to differentiate these overlapping emotional spaces, leading to frequent misclassifications. The confusion matrix further confirms this phenomenon by showing significant confusion between *Happy*, *Party*, and *Romantic* classes.

Overall, the five-category experiment demonstrates both the feasibility and limitations of instrumental MER, highlighting emotion ambiguity as a fundamental challenge.

Table 5.2. Performance metrics for four-category instrumental MER (Happy removed)

Emotion	Precision	Recall	F1-score	Accuracy (%)
Devotional	0.69	0.73	0.71	73.33
Party	0.69	0.73	0.71	73.33
Romantic	0.73	0.53	0.62	53.33
Sad	0.62	0.71	0.67	71.43
Overall	—	—	**0.68**	67.80

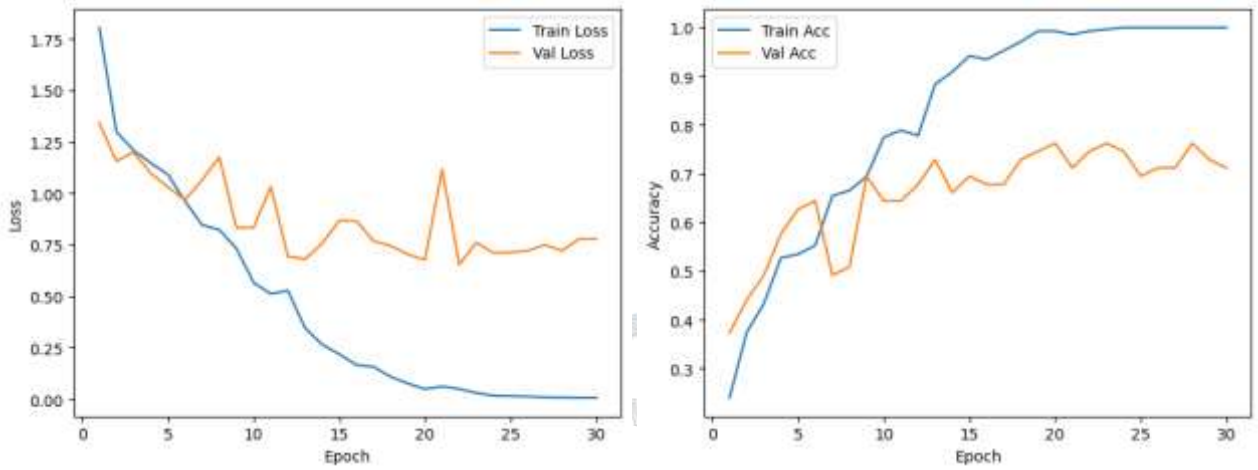

Figure 5.2 (a) Training and validation loss curves (b) Training and validation accuracy curves for 4-category the CNN-based instrumental MER model.

We investigate the impact of emotion ambiguity on classification performance, a second experiment is conducted by excluding the *Happy* emotion class. The CNN model is retrained on the remaining four categories: *Devotional*, *Party*, *Romantic*, and *Sad*. This experiment serves as an **ablation study** to evaluate how emotion taxonomy design influences instrumental MER.
The four-category configuration yields a test accuracy of **67.80%** and a macro-averaged **F1-score of 0.68**, representing a clear improvement over the five-category setting. The reduced gap between training and validation performance indicates improved generalization, suggesting that removing highly overlapping classes simplifies the decision boundaries learned by the CNN.
From a theoretical standpoint, this improvement can be explained by reduced class overlap in the spectral–temporal domain. *Party* music is primarily characterized by strong rhythmic patterns and high energy, *Sad* music by slower tempo and low spectral energy, and *Devotional* music by repetitive and harmonically stable structures. These distinctions allow convolutional filters to learn more separable feature representations. Class-wise results further support this analysis. *Devotional* and *Party* emotions achieve accuracies of **73.33%**, while *Sad* maintains **71.43%** accuracy. However, *Romantic* emotion shows comparatively lower performance (**53.33%**), which can be theoretically attributed to its acoustic proximity to *Sad* emotion. Both categories often share moderate tempo, smooth melodic contours, and softer timbral characteristics, resulting in partial feature overlap even after removing the *Happy* class. This experiment confirms that emotion class selection plays a critical role in instrumental MER and that reducing ambiguity leads to measurable performance gains.
A three-fold cross-validation experiment is conducted on the four-category dataset. The obtained fold-wise accuracies are **67.18%**, **53.44%**, and **71.76%**, resulting in a **mean accuracy of 64.12%**. Although performance varies across folds due to limited dataset size, the average accuracy remains consistent with the single-split evaluation. These findings demonstrate that the proposed CNN model generalizes reasonably well across different data partitions and does not rely on a specific train–test split.

The experimental results demonstrate that CNN-based deep learning models can effectively capture emotional characteristics from **instrumental Hindi music**. The consistent improvement observed after removing the *Happy* emotion class highlights the challenges associated with emotion ambiguity in instrumental MER. The cross-validation analysis further emphasizes the importance of robust evaluation strategies when working with culturally rich but limited datasets. Overall, the results confirm that instrumental music contains meaningful emotional cues, while also revealing the limitations of single-modality approaches in distinguishing closely related emotions.

V. CONCLUSION

This paper presented a comprehensive CNN-based deep learning framework for instrumental Hindi music emotion recognition, with a specific focus on analyzing emotions conveyed without vocal or lyrical information. By converting voice-removed music signals into spectrogram representations and employing a convolutional neural network, the proposed system effectively learned hierarchical spectral–temporal features associated with musical emotions.The study conducted multiple experimental evaluations, including five-category classification, four-category classification, and three-fold cross-validation. The results demonstrate that instrumental music alone contains sufficient emotional cues to achieve meaningful classification

performance, with notable improvements observed after removing highly ambiguous emotion classes. The cross-validation results further validate the robustness and generalization capability of the proposed approach. From a broader perspective, this work contributes to language-specific and culturally diverse MER research by addressing the limitations of Western-centric datasets and lyric-dependent approaches. The findings emphasize the importance of emotion taxonomy design, dataset characteristics, and evaluation methodology in music emotion recognition, particularly for instrumental and traditional music.

Future work will focus on expanding the dataset size and diversity to improve generalization performance. Incorporating vocal and lyrical information within a multimodal learning framework represents a promising direction for reducing emotion ambiguity. Additionally, exploring advanced architectures such as CNN–LSTM hybrids, attention-based models, and self-supervised pre-trained music representations may further enhance emotion recognition performance in complex and culturally rich music datasets.

REFERENCES

[1] Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A., and Turnbull, D. 2010. Music emotion recognition: A state of the art review. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2010: 255–266.

[2] Yang, Y.-H., and Chen, H. H. 2012. Machine recognition of music emotion. *ACM Transactions on Intelligent Systems and Technology*, 3(3): 1–30.

[3] Yang, X., Dong, Y., and Li, J. 2017. Review of data features-based music emotion recognition methods. *Multimedia Systems*, 24(4): 365–389.

[4] Han, D., Kong, Y., Han, J., and Wang, G. 2022. A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6): 1–25.

[5] Moysis, L., Iliadis, L. A., Sotiroudis, S. P., Boursianis, A. D., Papadopoulou, M. S., Kokkinidis, K.-I. D., Volos, C., Sarigiannidis, P., Nikolaidis, S., and Goudos, S. K. 2023. Music deep learning: Deep learning methods for music signal processing—A review of the state-of-the-art. *IEEE Access*, 11: 17031–17070.

[6] Kang, J., and Herremans, D. 2025. Towards unified music emotion recognition across dimensional and categorical models. *arXiv preprint*, arXiv:2502.03979.

[7] Liu, A. H., Chi, P.-H., Hsu, W.-N., Lee, H.-Y., and Chen, H.-T. 2021. Self-supervised learning of representations for music understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 176–189.

[8] Yang, Y., Dong, Y., and Li, J. 2020. Multimodal music emotion recognition using audio and lyrics. *IEEE Transactions on Affective Computing*, 11(3): 1–14.

[9] Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., and Patras, I. 2012. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1): 18–31.

[10] Sharma, R., Verma, A., and Gupta, P. 2024. Deep learning-based music emotion recognition on the MER500 dataset using angular margin loss. *International Journal of Multimedia Information Retrieval*, 13(2): 1–14.

[11] Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., and Lamere, P. 2011. The million song dataset. *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2011: 591–596.

[12] Pandey, B., and Mishra, P. 2019. Raga-based emotion modeling in Indian classical music. *Journal of New Music Research*, 48(4): 345–358.

[13] Chordia, P., and Rae, A. 2008. Understanding emotion in raagas: An empirical study of rasa theory. *Proceedings of the International Conference on Music Perception and Cognition*, 2008: 110–115.

[14] Gulati, S., Serra, X., Ganguli, K. K., and Rao, P. 2016. Timbre analysis of Indian musical instruments. *Journal of the Acoustical Society of America*, 140(3): 1567–1581.

[15] Biswas, A., and Dey, N. 2015. Clustering-based analysis of Hindi music using audio features. *International Journal of Signal Processing Systems*, 3(2): 145–151.

[16] Rao, P., Ross, J. C., Ganguli, K. K., Ishwar, V., Bellur, A., and Murthy, H. A. 2019. Classification of Indian classical music using convolutional neural networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019: 736–740.

[17] Sharma, M., and Kumar, A. 2022. Music emotion recognition in Hindi popular songs using convolutional neural networks. *Multimedia Tools and Applications*, 81(12): 16845–16862.

[18] Yang, S., Zhao, Y., and Li, Q. 2018. JUMusEmoDB: A music emotion database for instrumental emotion recognition. *Proceedings of the International Conference on Multimedia Modeling*, 2018: 500–512.

[19] Moysis, L., and Goudos, S. K. 2023. Traditional music deep learning: A comprehensive survey. *IEEE Access*, 11: 44567–44590.

[20] Oppenheim, A. V., and Schafer, R. W. 2010. *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Pearson Education.

[21] Rabiner, L., and Schafer, R. W. 2011. *Theory and Applications of Digital Speech Processing*. Upper Saddle River, NJ: Pearson Education.

[22] Choi, K., Fazekas, G., Sandler, M., and Cho, K. 2016. Convolutional recurrent neural networks for music classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016: 2392–2396.

[23] LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature*, 521(7553): 436–444.

[24] Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25: 1097–1105.

[25] Pons, S., and Serra, X. 2017. Designing efficient convolutional neural networks for music classification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017: 331–335**.**

[26] Kingma, D. P., and Ba, J. 2015. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*, 2015: 1–15.

[27] Goodfellow, I., Bengio, Y., and Courville, A. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

[28] Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1995: 1137–1143.
[29] Arlot, S., and Celisse, A. 2010. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4: 40–79.
[30] Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874.
[31] Powers, D. M. W. 2011. Evaluation: From precision, recall and F-measure to ROC. *Journal of Machine Learning Technologies*, 2(1): 37–63.
[32] Velankar, M., "MER500: A Hindi film music emotion dataset with 500 clips categorized as Romantic, Happy, Sad, Devotional, and Party," dataset (MER500), available via GitHub/Kaggle (accessed March 10, 2024).