Design and Implementation of a Machine Learning Based Hindi Music Emotion Classification System

Akanksha Gupta[1*] and Anamika Singh[2]

[1]LNCT University,Bhopal, India
[2]LNCT University,Bhopal, India
*akkuregister.90@gmail.com

Abstract. Music Emotion Recognition plays an essential part in detecting particular emotion classes in songs by analysing their emotional content. In this study, we design and implement a MER system based on machine learning using acoustic feature analysis on the Hindi Music dataset MER500. Two experiments are conducted using different window and hop size configurations, i.e. 2048×1024 and 1024×512, to analyse the effect of temporal segmentation on emotion classification performance. From each configuration, a set of relevant audio features, including MFCCs, spectral descriptors, chroma features, energy, and tempo-related attributes, are extracted. These features are then classified using various machine learning algorithms such as Support Vector Machine, Random Forest, K-Nearest Neighbours. The results from the experiment demonstrate that window–hop size selection significantly influences emotion classification accuracy. Among the tested configurations, a balanced time–frequency resolution provides superior performance and computational efficiency. The proposed system offers an effective and scalable solution for Hindi music emotion recognition and highlights the importance of optimised signal processing parameters for improved classification performance.

Keywords: Music Emotion Recognition, Machine Learning, Audio Features, Music Signal Processing.

1 Introduction

Music significantly influences human emotions. It is critical to identify the emotion from a piece of music. This emotional information is used in the retrieval and recommendation of music. Every piece of music conveys some feeling in the human [1]. In recent years, much research has been done in recognising music emotions automatically using various machine and deep learning algorithms.Music and emotion arestrongly related to each other as every song is associated with an emotion. Music is the art form where sound is used to convey feelings. Music can evoke, enhance, and regulate emotions that motivate researchersto do more research in the field. MER has many applications, like music recommendation [2], music retrieval [3], generatinganimated imagery based on a pieceof music, psychotherapy and so on. It pertains

S. Bhalerao et al. (eds.), *Proceedings of the 2nd International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2026)*, Advances in Intelligent Systems Research 207,
https://doi.org/10.2991/978-94-6239-678-4_12

to the interdisciplinary domainsof music psychology, audio signal processing, and natural language processing(NLP).

Considerable work has been done in categorizing music into genres, moods,and instrumentation. But, classification of music on the basis of emotion demands great attention, as this field has many challenges, subjectivity of emotion, difficulty of emotion annotation, selection of music features and choosing aclassification algorithm is another challenge.

In recent years, machine learning methodologies have been extensively utilized for Music Emotion Recognition (MER) owing to their capacity to discern intricate patterns from audio characteristics. A typical MER system involves signal preprocessing, feature extraction, feature selection, and classification. Acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral descriptors, chroma features, energy, and tempo-related attributes are commonlyutilized to signify the emotional essence of music. The efficacy of these featuresis significantly dependent upon the time-frequency resolution of the signal representation, which is determined by parameters such as window size and hopsize during spectrogram generation. Despite the importance of these parameters, limited studies have systematically analyzed their effect on the performance ofmachine learning based MER systems, especially for Hindi music.

This paper is organized as follows: Section 2 gives a review of the related workon music emotion recognition and machine learning based approaches. Section 3describes the proposed system architecture, including dataset description, signalpreprocessing, extraction of features, and classification methods. Section 4 explains the experimental setup and the different window–hop size configurationsused for performance evaluation. Section 5 discusses the experimental results andprovides a comparison of various machine learning models. The work is finallyconcluded, and future research possibilities are outlined in Section 6.

2 Past Work

A Music emotion recognition system requires the emotion-labelled data. There are two approaches, i.e. categorical and dimensional. Hence, two types of models are popular in MER: one is a categorical model and the second is a dimensional model [12]. categorical models (such as Hevner model) [8] classify the emotion into discrete emotion categories like sad, happy, fear and anger. Dimensional model(Russell's V-A model[9], Thayer model[10]), A point in the two-dimensional emotion space of valence and arousal is used to represent emotion. Valence axis is related to the pleasantness or affective state of a song. positive valence indicates a pleasant song, whereas negative valence indicates an unpleasant song.The arousal axis is associated with the

energy of a song. positive arousal indicates high energy, like excitement, whereas negative arousal indicates low energy, like a sleepy state. Choice of emotion model is up for debate. On the basis of the emotion model, MER computation models are of two types: classifier models and regression models. A classifier-based approach predicts the discrete emotion output. In regression-based approach [13], model predict the Valance and arousal values for the music clip. MER is still a topic of ongoing study. The most prestigious worldwide audio recovery and analysis competitions, Music Information Retrieval Evaluation eXchange (MIREX), demonstrated the importance of MER in 2007 by including audio mood categorisation (AMC) [4] in its competition tasks.

Kim et al. [5] conduct a thorough survey that emphasises content-based and context-based information for the MER method. Kim et al have not involved the ground truth data that was later involved by yang et al [6]. The perceived emotions of people are reflected in ground truth data. [6] gives MER methods based on three different feature data, only audio feature, only ground truth data, and a combination of both. ground truth data are of two types, first is label type, and the second is numerical type data.

Panda et al [11][14] investigate the audio features that are relevant to the emotion, including low-level features(energy, zero crossing rate, spectral feature,MFCC), perceptual (rhythm clarity, modality) and high-level features (Dynamics Variation, genre, danceability). [14] studied the relationship between different 8 musical dimensions (rhythm, harmony, tone colour, dynamics, expressivity, texture and form) and specific emotion. There are many tools such as PsySound[15], MARSYAS[16] and MIR Toolbox[17] for extracting the various audio features, which are used in many classification methods to classify the emotion. In MER field, commonly used classification methods are support vector machine (SVM) [18] [19], k-nearest neighbor (KNN) [20][21], neural network (ANN) [22], radial basis function ANN (RBF-ANN)[23],Gaussian Mixture Model(GMM)[24][25], Random Forest [26][27].

Several publicly available datasets have been widely used to validate machine learning based MER systems. The DEAP Dataset [28] has been extensively utilized for emotion analysis using physiological signals as well as music stimuli. Researchers applied KNN, SVM, and RF classifiers on DEAP-based audio features and reported classification accuracy in the range of 70–85%. Another commonly used dataset is the EMO-DB[29], on which multiple ML models including SVM and Gaussian Mixture Models (GMM) have been evaluated for emotion recognition. Although EMO-DB is a speech dataset, its classification results have influenced MER system design and feature selection strategies.

The DEAM Dataset[30] is another widely used music emotion dataset containing dynamic valence and arousal annotations. Machine learning based regression and classification models such as SVM and Random Forest have beenapplied to DEAM using MFCC, spectral contrast, and chroma features, achieving notable prediction

performance. Panda et al.[15] proposed novel audio features by combining harmonic, spectral, and temporal representations for MER and evaluated them using SVM and KNN classifiers on benchmark datasets, reporting improved classification accuracy over baseline features. Their work highlighted the importance of feature-level optimization in machine learning driven MER systems.

In the Indian music domain, relatively limited studies have been reported. Moreover, although deep approaches such as Convolutional Neural Networks(CNNs) [31] have recently shown strong performance in large-scale MER systems, they demand extensive labeled datasets and significant computational resources. In contrast, traditional machine learning based systems remain highly effective and computationally efficient for medium-sized datasets such as the MER500 Hindi Music Dataset[7], where well-designed handcrafted features still outperform complex deep learning models in constrained scenarios. Although machine learning based MER has been widely studied across different datasets,a systematic multi-experimental evaluation of window and hop size variations for Hindi music emotion recognition remains untouched. This work aims to bridge this gap by analysing the effect of different time–frequency resolutions under a unified machine learning framework.

3 Research Methodology

The overall method used in the development and deployment of the suggested machine learning-based Hindi MER system is explained in this part. The complete workflow of the system consists of 1. dataset preparation, 2. pre-processing of signal, 3. extraction of feature, 4. multi-experiment configuration using different window–hop sizes, machine learning based classification, and performance evaluation.

3.1 Dataset

In this work, the MER500 dataset is used. This dataset has five emotion classes:devotional, happy party, romantic and sad. Each emotion class has 100 songsin .wav format. The size of each song is 10 sec. A exact standard format ispreprocessed for the music database, including accuracy (16 bits) and samplingfrequency (44,100 Hz).

3.2 Preprocessing

All audio signals from the MER500 Hindi Music Dataset are first converted into a uniform waveform format and resampled to a fixed sampling rate to maintain consistency across the dataset. Amplitude normalisation is applied to reducevariations in signal intensity, and silent or low-energy segments are removedto preserve only emotion-relevant musical information. Then, we perform framebased signal analysis directly on the time-domain audio signal. The continuousaudio waveform is divided

into short overlapping frames using fixed frame lengths and hop sizes. To analyze the effect of temporal resolution on emotion classification performance, three different frame–hop configurations are employed. In the first configuration, uses a frame size of 2048 samples and a hop size of 1024 samples (2048x1024)., corresponding to 50% overlap between adjacent frames. The second setup used a hop size of 512 samples and a frame size of 1024 samples (1024x512).

3.3 Feature Extraction

In this work, a total of 46 handcrafted acoustic features [14] are extracted from each audio signal to characterize its emotional content. Feature extraction is performed on short-time frames obtained using the framing strategy described in preprocessing Section, where each frame is multiplied by a Hann window before analysis. For every frame, temporal, spectral, cepstral, and chroma-based descriptors are computed, and then aggregated at the clip level to obtain a fixed-length feature vector.

– First, two temporal features are retrieved directly from each time-domainframe. The short-time energy is computed as the sum of squared amplitude samples within a frame, which reflects the local signal energy and correlateswith perceived loudness and intensity variations. The zero-crossing rate (ZCR) is calculated as the average rate at which the signal changes sign,providing a measure of signal noisiness and the presence of high-frequencycomponents.

– Second, a set of six spectral features is derived from each frame's magnitude spectrum, obtained using a 512-point Fast Fourier Transform (FFT).The spectral centroid represents the "centre of mass" of the spectrumand indicates the sound brightness. The spectral roll-off is defined as thefrequency below which a fixed percentage (85%) of the total spectral energyis contained. The spectral flatness measures the degree of noise-like ortone-like the spectrum is by taking the ratio of geometric mean to arithmetic mean of the magnitude spectrum. The spectral crest factor captures the peakiness of the spectrum by comparing the maximum magnitudeto the average magnitude. The spectral flux measures the frame-to-framechange in the spectrum by computing the Euclidean distance between consecutive magnitude spectra, and the spectral bandwidth is calculated asthe second-order moment of the spectrum around the centroid, reflecting thespread of spectral energy.

– Third, To capture perceptually motivated timbral characteristics, Mel Frequency Cepstral Coefficients (MFCCs) are calculated. For each frame,the magnitude spectrum is mapped onto a 26-channel Mel filterbank, thelog energies of the Mel bands are evaluate, and a Discrete Cosine Transform is applied to obtain 13 MFCCs. In addition to static MFCCs, theirdynamic behaviour is modelled using delta MFCCs, which are computed viaregression over a ±2-frame temporal window, resulting in another set of 13delta coefficients. Together, the MFCC and delta MFCC features providethe spectral envelope and its temporal evolution.

– Fourth, Harmonic and pitch-class information is incorporated through 12 chroma features, which project the spectral energy onto 12 pitch classes (semitones) of the musical octave, relative to a reference frequency (A4 =440 Hz). For each frame, spectral magnitudes are accumulated into these 12 chroma bins, and the resulting chroma vector is normalized to form a probability-like distribution. These features are particularly useful for distinguishing emotions associated with different harmonic and melodic structures.

After extracting the features from each frame, their average and standard deviation are computed to generate the feature vector.

3.4 Machine Learning Classifier

In this study, three different classifiers are implemented namely Support Vector Machine (SVM), Random Forest (RF) and K-Nearest Neighbors (KNN). The feature vector derived from the feature extraction process is given to these classifiers for emotion recognition. The SVM classifier by maximising the margin between class boundaries, is used with a non-linear kernel to divide emotion classes in a high-dimensional feature space The K-Nearest Neighbors (KNN) classifier uses the majority vote of a test sample's k nearest neighbours to decide the class label using Euclidean distance as the similarity metric. Whereas in the Random Forest (RF) classifier, A number of decision trees are trained using samples and feature subsets selected at random, which is resistant to overfitting and noise because the majority vote determines the final prediction. It is make up an ensemble learning technique.

4 Experimental Setup

For all experiments, a python programming environment is used for togetherwith common learning libraries such as NumPy, Pandas, and scikit-learn. Tomaintain the initial class distribution across all five emotion categories, the entire dataset is split into 80:20% training:testing respectively groups using stratified sampling. All feature vectors are normalised using z-score standardisation,to make each feature has zero mean and unit variance. Standard multi-classclassification measures, including as accuracy, precision, recall, and F1-score, aswell as the confusion matrix for class-wise error analysis, are used to assess eachclassifier's performance.

5 Result

The performance evaluation of the suggested machine learning-based Hindi Music Emotion Recognition (MER) system on the MER500 Hindi Music Dataset under

various frame-hop size settings is shown in this part. Precision(P), recall(R), F1-score(F), and average accuracy are used to evaluate the effectiveness of SVM, RF, and KNN classifiers for five emotion classes: Devotional, Happy, Party, Romantic, and Sad.

5.1 Experiment 1

Table 1 summarizes the performance comparison of the three classifiers usingthe frame size and hop size 2048 and 1024, respectively. The Random Forest(RF) classifier achieved the highest average accuracy of 71.72%, outperformingSVM (67.68%) anKNN (55.56%). Class-wise analysis shows that RF performsparticularly well for the Devotional (F1 = 0.84) and Party (F1 = 0.82) classes,indicating its robustness in handling energetic and rhythm-dominated emotions.The SVM classifier shows

Table 1 Experiment 1 Comparison of Different Classifiers Across Five Emotion Classes with MER500 dataset with frame size and hop size (2048x1024)

	SVM			KNN			RF		
Class 5_CAT	P	R	F	P	R	F	P	R	F
Devotional	1	1	1	1	1	1	1	1	1
Happy	1	1	1	1	1	1	1	1	1
Party	1	1	1	1	1	1	1	1	1
Romantic	0	0	0	0	1	0	1	1	1
Sad	1	1	1	1	1	1	1	1	1
Avg. Acc.	67.68%			55.56%			71.72%		

Table 2 Experiment 2 Comparison of different Classifiers Across Five Emotion Classes with MER500 dataset with frame size and hop size (1024x512)

Class	SVM			KNN			RF		
5_CAT	P	R	F	P	R	F	P	R	F
Devotional	1	0.8	0.9	0.7	0.7	0.7	0.8	1	0.9
Happy	0.6	0.7	0.7	0.5	0.7	0.6	0.8	0.6	0.7
Party	0.9	0.7	0.7	0.8	0.6	0.7	0.7	0.9	0.8
Romantic	0.4	0.6	0.5	0.4	0.5	0.4	0.6	0.5	0.5
Sad	0.8	0.8	0.8	0.7	0.6	0.6	0.8	0.8	0.8
Avg. Acc.	72.73%			58.59%			73.74%		

competitive performance for the Party class (F1 =0.84), but its performance drops for the Romantic class (F1 = 0.41), suggestingdifficulty in modeling subtle emotional patterns under this framing configuration. The KNN classifier records the lowest

overall performance, with an averageaccuracy of only 55.56%, mainly due to confusion among acoustically similaremotion classes such as Romantic and Sad. Overall, the results indicate thatthe 2048×1024 configuration favors ensemble-based learning (RF) over distancebased learning (KNN).

5.2 Experiment 2

Table 2 reports the results obtained using a smaller frame size and hop sizeof 1024×512, which provides a higher temporal resolution. Under this configuration, the performance of all classifiers shows a noticeable improvement. TheRandom Forest classifier again achieved the best average accuracy of 73.74%,followed by SVM with 72.73%, while KNN obtained 58.59.

The improvement in SVM performance compared to Experiment–1 indicates that smaller frame and hop sizes better capture short-term emotional variations, which are important for emotions such as Devotional (SVM F1 = 0.88) and Sad (SVM F1 = 0.80). RF continues to demonstrate consistent performance across multiple classes, achieving strong results for Party (F1 = 0.80) and Sad (F1 =0.78). KNN, although improved, still lags behind due to its sensitivity to high- dimensional feature distributions and overlapping emotional clusters.

6 Conclusion

This work presents the design and deployment of a machine learning-based Hindi Music Emotion Recognition (MER) system utilizing a frame-based audio signal analysis methodology. A thorough collection of 46 precisely produced acoustic features was derived from each frame. The extracted features were evaluated using three widely used classifiers, namely Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF).The experimental findings clearly indicate that the selection of window and hop size is important to emotion recognition efficiency. The 1024x512 framing combination regularly surpassed the larger framing setup, demonstrating that finer temporal segmentation is more adept at capturing short-term emotional fluctuations in Hindi music. The Random Forest classifier is the best model for the suggested MER system since it performed the best and most consistently in every experiment. The results of this work verify that classifier selection and temporal segmentation approach have a major impact on MER performance, and that improving these design
parameters can result in appreciable increases in classification accuracy.

References

1. Feng, Y., Zhuang, Y., Pan, Y.: Popular music retrieval by detecting mood. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 375–376. ACM, New York (2003)
2. Cheng Z Y, Shen J L, Zhu L, Kankanhalli M,Nie L Q. Exploiting music play sequence for music recommendation. In: Proceedings of the26th International Joint Conference on Artificial Intelligence. 2017,pp. 3654-3660.
3. Cheng Z Y, Shen J L, Nie L Q, Chua T S, Kankanhalli M. Exploring user-specific information in music retrieval. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017,pp.655–664.
4. Hu X. , Downie J. , Laurier C. , Bay M. , and Ehmann A. , "The 2007 MIREX
audio mood classification task: Lessons learned," in Proc. of the Intl. Conf. on
Music Information Retrieval, Philadelphia, PA, 2008.
5. Kim Y E, Schmidt E M, Migneco R, Morton B G, Richardson P, Scott J, Speck J A, Turnbull D. Music emotion recognition: a state of the art review. In: Proceedings of the 11th International Society for Music Information Retrieval Conference. 2010, pp 255–266
6. Yang X Y, Dong Y Z, Li J. Review of data features-based music emotion recognition methods. Multimedia System, 2018, 24(4): 365–389
7. M. Velankar, "MER500: A Hindi film music emotion dataset with 500 clips categorised as Romantic, Happy, Sad, Devotional, and Party," dataset (MER500), available via GitHub/Kaggle (accessed March 10, 2024).
8. Hevner K. , "Experimental studies of the elements of expression in music," American Journal of Psychology, vol. 48, no. 2, pp. 246–267, 1936.
9. Posner J, Russell J A, Peterson B S. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychology. Development and Psychopathology, 2005, 17(3): 715–734
10. Thayer R. The Biopsychology of Mood and Arousal. 1st ed. Oxford: Oxford University Press, 1989
11. Panda R., Malheiro RM, Paiva RP (2018) Novel audio features for music emotion recoginition. IEEE Transactions on affective computing.
12. Yang Y H, Chen H H (2012) Machine recogination of music emotion: a review. ACM Trans Intell Syst Technol 3(3): 40: 1-4-:30
13. Yang Y.H., Lin Y.C., Su Y.F., Chen H.H.(2007) Music emotion classification: a regression approach. In: International conference on multimedia and expo, pp 208-211
14. Panda R., Malheiro RM, Paiva RP (2020) Audio features for music emotion recognition: a survey. IEEE Transactions on affective computing.
15. Cabrera D. et al (1999) Psysound: a computer program for psychoacoustical analysis.In: Australian Acoustical Society conference. Vol 24, pp 47-54.

16. Tzanetakis G, Cook P (1999) Marsyas: a framework for audio analysis. OrganisedSound 4(3): pp 169-175.
17. MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio.January2007 Conference: Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007
18. Chin, Y.H. Lin, P.C. Siahaan, E., Wang, I.C. Wang, J.C.:Music emotion classification using double-layer support vector machine. In 2013 International Conference on Orange Technology (ICOYT), pp- 193-196. IEEE, New York (2013)
19. Han B.J. Rho S. Jan S. Hawng E. (2010) Music emotion classification and contextbased music recommendation multimed tool Appl 47 (3): 433-460
20. Dewi, K.C., Harjoko, A.: kId's song classification based on mood parameters usingk-nearest neighbour classification method and self organing map. In: 2010 InternationalConference on Distributed framework and application(DFmA), pp 1-5, IEEE, newyork(2010)
21. Pao T.L., Cheng Y.M., Yeh J.H., Chen Y.t.Pai, C.Y., Tsai Y.W.: Comparisonbetween weighted d-knn and other classifier for music emotion recognition. IN3rd internation conference on Innovative computing information and control 2008.ICICIC'08, pp530-530,IEEE, new york(2008)
22. Feng Y., Zhuang Y., Pan Y.,: Popular music retreival by detecting mood. In : Proceedings of the 26th annual International ACM SIGIR conference on research anddevelopment in information retrieval, pp 375-376, ACM, new york 2003
23. Ooi C S, Seng KP, Ang LM , Chew LW(2014) A new approach of audio emotionrecognition, Expert Syst Appl 41(13), 5858-5869
24. Lu L,Liu D, Zhang H (2006) Automatic mood detection and tracking of musicaudio signal.IEEE tans Audio speech lang process 14(1),5-18
25. Zao L, Cavalcante D, Coelho R (2014) Time frequency feature and ams-gmm mask for acoustic emotion classification, IEEE signal process Lett 21(5): 620-624
26. Zhang F, Meng H, Li M(2016) Emotion extraction and recognition from music.In: International conference on natural computation, fussy system and knowledgediscovery pp 1728-2128.
27. Boser, B., Guyon, I., Vapnik, V.(1992): A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh
28. Koelstra S. et al., "DEAP: A Database for Emotion Analysis Using Physiological Signals," IEEE Transactions on Affective Computing, 2012.
29. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B.,"A database of German emotional speech," Proceedings of Interspeech, Lisbon, Portugal, 2005, pp. 1517–1520.
30. Aljanaki, A., Yang, Y.-H., Soleymani, M., "Developing a benchmark for emotional analysis of music," Proceedings of the 25th ACM International Conference on Multimedia (ACM MM), 2017, pp. 508–516.

31. Liu, T., Han, L., Ma, L., Guo, D. Audio-based deep music emotion recognition. In Proceedings of AIP (Vol. 1967), May 2018.