Multimedia Tools and Applications (2022) 81:8853–8870
https://doi.org/10.1007/s11042-022-11975-4

Multi-label emotion recognition from Indian classical music using gradient descent SNN model

Bhavana Tiple[1] · Manasi Patwardhan[2]

Received: 1 September 2020 / Revised: 28 July 2021 / Accepted: 3 January 2022
Published online: 8 February 2022

Abstract
Music enthusiasts are growing exponentially and based on this, many songs are being introduced to the market and stored in signal music libraries. Due to this development emotion recognition model from music contents has received increasing attention in today's world. Of these technologies, a novel Music Emotion Recognition (MER) system is introduced to meet the ever-increasing demand for easy and efficient access to music information. Even though this system was well-developed it lacks in maintaining accuracy of the system and finds difficulty in predicting multi-label emotion type. To address these shortcomings, in this research article, a novel MER system is developed by interlinking the pre-processing, feature extraction and classification steps. Initially, pre-processing step is employed to convert larger audio files into smaller audio frames. Afterwards, music related temporal, spectral and energy features are extracted for those pre-processed frames which are subjected to the proposed gradient descent based Spiking Neural Network (SNN) classifier. While learning SNN, it is important to determine the optimal weight values for reducing the training error so that gradient descent optimization approach is adopted. To prove the effectiveness of proposed research, proposed model is compared with conventional classification algorithms. The proposed methodology was experimentally tested using various evaluation metrics and it achieves 94.55% accuracy. Hence the proposed methodology attains a good accuracy measure and outperforms well than other algorithms.

Keywords Convolutional neural network · Spiking neural network · Gradient descent · Temporal · Spectral · Short Term Fourier Transform

✉ Bhavana Tiple
bhavana.tiple@mitwpu.edu.in

[1] School Of SCET, Dr. Vishwanath Karad MIT World Peace University, Pune, India

[2] TCS Innovation Labs, Pune, India

1 Introduction

People listen to music in accordance with their moods, and music also change the emotions of humans [19]. Emotions are key to people's feelings and thoughts. Thus, it become prominent to categorize music according to the type of emotion it expresses. Researchers claimed that music can resonate with our nerve tissue. Based on the research it was observed that young children who do not know what music means, regardless of language and what they see, respond to what they hear. It is a biological nature that interacts between the rhythm and their brain [5, 20]. For a long time, many researchers have explored the relationship between music and emotion [17]. Music can evoke more than one different emotion at the same time which means a piece of music may simultaneously belong to more than one class. Single-label classification does not model this multiplicity. Hence the authors in [4, 32] gave a detailed introduction about MER and described MER as a multi-label classification problem. In multi-label classification, each series is considered a characteristic with multiple labels, and each label represents one type of emotion. Also, it is important to select the features for the classification of a task. Without proper feature extraction stage, it is not easy to extract emotion from classical music. Acoustic feature selection approaches are more prevalent in other feature selection approaches to music representation [28]. Features such as rhythmic, spectral [29] and timbre are included in the acoustic features. Spectral flux, spectral flatness measurement, spectral crest factor, spectral centroid and spectral roll of belongs to spectral features. Using these features the entire music is going to convert into digital form [18].

Even after extracting the musical features, it is difficult to distinguish the mood from the Indian classical music (ICM), as it has only two parts and they are raga and rasa [2, 14]. The seven basic of ICM are Swaras (notes) combinations are called ragas, namely [Sa (Shadja), Ri (Rishabha), Ga (Gandhara), Ma (Madhyama), Pa (Panchama), Dha (Dhaivata), Ni (Nishada)]. Each raga defines specific notes and split up into rasas based on certain rules. There are nine rasaas in the ICM and they are: 1) Shringar (Love), 2)Hasya(Humor), 3) Karuna(Pathos), 4) Rudra (Anger), 5) Veer (Heroism), 6) Bhayanaka(Terror), 7) Vibhatsa (Disgust), 8) Adbhuta (Wonder), 9)Shanta (Calm). In the ICM there is a raga-rasa theory which induces a specific emotion in the mind of the listener. Each performance of a given raga in ICM is supposed to build a common mood/rasa amongst the listeners. But extracting exact emotion is quite complex task because the emotions or mood created by music for different or same listeners may vary at different times. So particularly in ICM, mood or emotion recognition poses an open issue. Therefore, choosing essential features is an essential step for emotion identification. This type of features are extracted and subjected into machine learning algorithms for the automatic detection of emotion from music notes [10].

Nowadays, choosing a classifier is also an important task of MER [23]. Some of the classifiers are calibrated label ranking classifiers which use support vector machine (CLRSVM) [15], Random k-label sets (RAKEL), binary relevance KNN (BRKNN), Multi-label k-nearest neighbour (MLKNN) and Back-propagation for multi-label learning (BPMILL) etc. Generally, the neurons in CNN and in DBN are classified by different activations like continuous valued, single and static. Still, discrete spikes are used for computing and transmitting information, by biological neurons. The spike time and their rate matters lot in the process. Moreover many improvements have been made to this MER system, it still does not meet the demand for accuracy measurement. Therefore an efficient system is needed to classify music emotion with high accuracy according to the emotional range of people [30, 31]. To solve this problem we use the latest technology called SNN (Spiking Neural Network)

Architectures [22]. This technique does not use any simple artificial intelligence parameters that directly use the spectrogram of music. In the classification algorithm, spiking neurons are activated when given information, as they are a close representation of the concept on the human brain [27]. This classification algorithm is the current trendy approach which achieves notable performance in all domains and almost all the problems have better accuracy compared to other machine learning methods [1, 9, 16]. Therefore, SNN's are more convenient than CNN's, DBN's and other existing approaches specifically for emotion recognition based on music tones. The contributions of proposed research are illustrated as follows:

- To segment the larger audio files into smaller frames, pre-processing steps are carried out.
- To extract the essential features from the pre-processed output, feature extraction process is employed. This will also improve the system performance by extracting only the relevant features.
- To classify multi-label emotions efficiently, a novel optimization based SNN algorithm is proposed.
- The optimization procedure follows gradient descent algorithm whereas network weights are effectively chosen by means of training phase.

With these contributions, accuracy for the proposed method will be improved and automatically, training error gets reduced. The step by step procedure to achieve these contributions are listed below:

1. Pre-processing– Initially the input audio files are collected and fed into the pre-processing phase where larger audio files are segmented into smaller frames.
2. Feature extraction- This phase extracts seven different features such as Mel Frequency Cepstral Coefficient, Spectral Flux, Spectral Centroid, Spectral Roll-off, Zero Crossing Rate, Compactness and Root Mean Square for pre-processed output image. Those features are subjected to classification stage for further processing.
3. Classification of multi label emotion type- Here training and testing the model is done with the help of novel gradient descent based SNN model.

The manuscript is organized as follows: Section 2 details about the related works. Section 3 demonstrates the proposed methodology. Dataset description and experimental results are elaborated in subsequent Section. Conclusion is provided in the final section.

2 Literature review

Some of the existing work related to MER systems are elaborated in this section.

Panda et al. [25] have designed the music emotion recognition system by introducing an emotionally relevant audio features. They survived the several features and detailed a deep knowledge about those features. Musical texture and expressive methods were developed to expose musical impressions. They created a public dataset which comprising 900 set of audio clips placed in Russell's emotion quadrants. The system was compared with traditional baseline feature and validated with 20 repetitions of 10-fold cross-validation. The developed model achieved 76.4% of accuracy which is not upto the mark for recognizing emotion labels.

Costa et al. [12] have presented an automatic music genre classification model. The audio signals were converted to spectrograms. Then from the visual formation, its relavant features were extracted. This model was based on time-frequency representation whereas texture

features were extracted to form the music genre classification systems. They utilized dataset comprising 900 music pieces and was separated into 10 music genres. Further the classifierr model used was trained with those extracted texture features. But accuracy attained was 60% by the exhibited method which was low compared to other classifiers.

Bhatti et al. [7] have exhibited the human emotion recognition system from the input audio music's. They used ECG signals for analysis and for those signals 13 features were extracted. After the feature extraction process, the classification algorithm classifies the four different emotions. The authors showed that MLP approach attains a maximum accuracy when compared with other k-NN and SVM classifiers. The music geners collected are rock and hiphop music which explores sad and happy emotion type. It also possessed rap and metal music geners which evokes sad and angry emotions. The model was tested with three classifiers namely MLP, k-NN and SVM acheives 78.11%, 72.80%, and 75.62% respectively. The accuracy of this emotion recognition model decreases as the number of emotions to be classified increases.

Malheiro et al. [21] have also focused the music emotion recognition model. This methodology extracted different set of features which were structural, semantic and stylistic characteristics. They also created a ground truth dataset containing set of 180 song lyrics based on Russell's emotion model. The authors also conducted an experimental study with regression and classification models in quadrant, arousal and valence categories. The value obtaine by F-measure is 82.7 and 85.6% to 80.1, 88.3 and 90%, respectively for the taken three classification experiments. They also validated their system with 771 lyrical musics from AllMusic platform and for this music genere. The presented method attained 73.6% accuracy which is low as compared to other techniques.

Baniya and Lee [3] have presented an automatic emotion classification system under the principle of rough set (RS) theory. Here at first, various set of features were extracted representing harmony, rhythm, dynamics and spectral. From the obtained set of features, the parameters related to statisticalmeasures were considered as attriburtes. These set of attributes were organized by means of RS based approach. Based upon the working style of RSmodel, the superflous features were eliminated. This theory helps to findthe relation among the attributes from the generated rules. The overall classification accuracy obtained by this method was 72%.

Mo and Niu [24] have presented music signal analysis framework for emotion classification. The authors combined orthogonal matching pursuit, Gabor functions, and the Wigner distribution function. They invented OMPGW method that consists of three-level schemes: the low-level, the middle-level and the high-level schemes. In low-level schemes, an adaptive time-frequency decomposition of music signals was generated by interlinking orthogonal matching pursuit with Gabor functions. For signal analysis, this presented model acheives higher temporal and spatial resolution. The Wigner distribution function is performed to attain time frequency energy distribution in middle-level schemes from the low-level schemes.Finally the High-level schemes are utilized tomodel the audio features for emotion classification system. The classification used in this research is support vector machines which acquires 69.53 accuracy.

All the existing works lacks in obtaining a good accuracy rate. Therefore it is difficult and challenging to retrieve the music information along with the recognition of musical emotions in accurate manner. This leads to degrade the system performance. The lack of solutions for such drawbacks motived us to do research in this field.

3 Research methodology

Humans express tons of emotions such as happy, anger, sad, surprise, Excitement etc. while listening music. But these emotions may vary by individuals for the same song. Moreover, the emotions associated with music are very subjective, so it is difficult to measure emotion based music accurately. By keeping this in mind a novel emotion recognition system is developed based on music interest. The layout of the proposed methodology is given in the below block-diagram.

Outline of the research

- Pre-processing via Short Term Fourier Transform (STFT).
- Temporal, Spectral and energy related feature extraction approaches.
- Music based Emotion recognition system using proposed classification algorithm.

For designing such a system, it is not easy to process the larger raga or audio file directly into it. Therefore, initially, pre-processing steps are done to segment the input files into smaller clip by using STFT. Afterwards, for those segmented audio clips, music-based features such as temporal, spectral and energy are extracted. At last, the extracted features are subjected to the proposed Gradient Descent based SNN classifier whereas the network weight get optimized by gradient descent algorithm so that the training errors are minimized.

3.1 Pre-processing

Generally, the input audio dataset is in lengthy format and is not directly suitable for classification process. Regarding this larger length of the music, pre-processing is an unavoidable step in this research methodology. Initially, the input music audio signal is segmented to 10 milliseconds audio framing blocks with the help of STFT technique. It describes the evolution of frequency contents of signals over time. The steps involved in STFT approach are given as follows:

1. Initially chose the window function with finite length.
2. Keep the window at t=10 on top of the signal.
3. Then the signal is truncated or trimmed by utilizing the window size.
4. For that trimmed signal, compute the Fourier transform and generate the outcome.
5. Slide the window to right in incremental way.
6. Repeat from step 3 to step 5, until window reaches the signal end.

The mathematically formula to process STFT technique is given in Eq. 1:

$$STFT\{f(s)\}(\tau,\omega) = \int_{-\infty}^{\infty} [f(s) \cdot w(s-\tau)] \cdot e^{-i\omega s} ds \tag{1}$$

Where $f(s) \cdot w(s-\tau)$ is the phase and magnitude of the signal over time and frequency. Here time axis and frequency axis of audio signal are termed as τ and ω respectively. $w(\tau)$ represents the window function centred around zero, $f(s)$ is the signal to be transformed i.e., difference between window function w and frequency ω. For each frame STFT of $f(s)$ is calculated. By which the spectral form of frames are generated. Following these steps, spectral form of frames are generated for the full audio clip. Thus, larger raga is segmented into smaller rasa's by managing the wave file into uniform format [11] (Fig. 1).

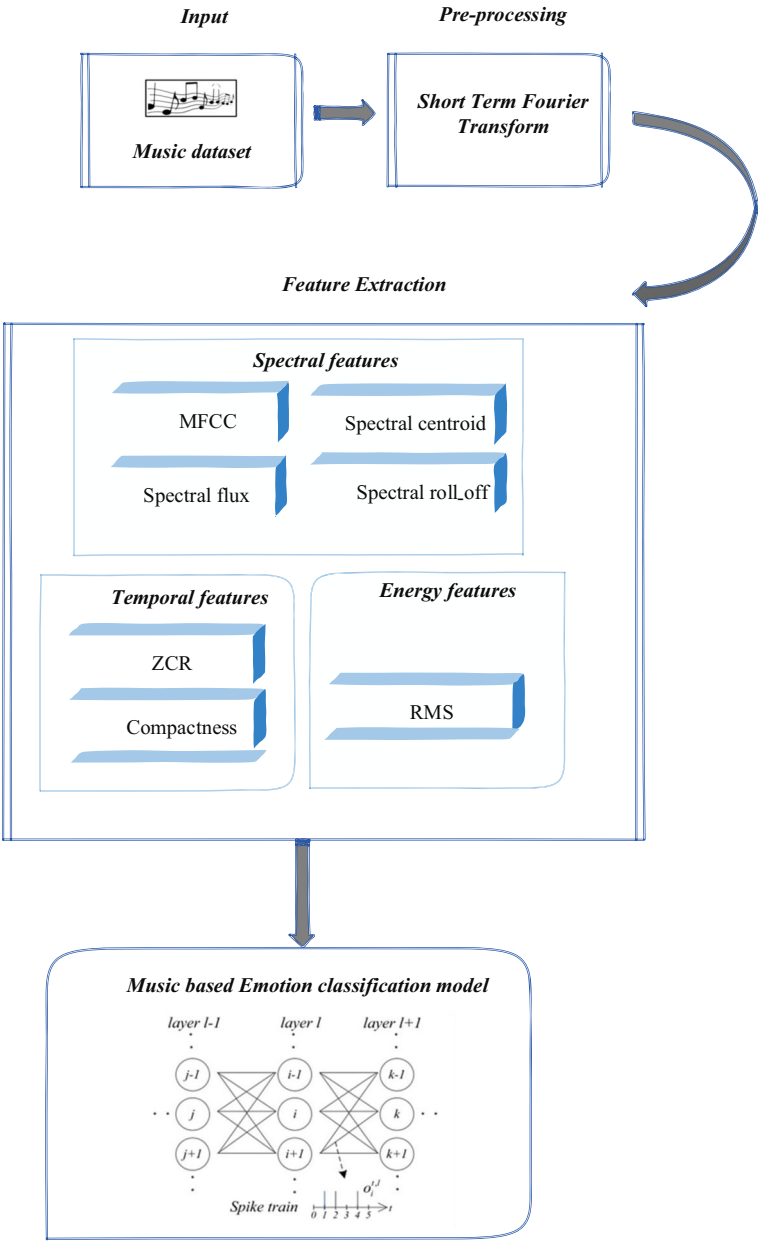

Fig. 1 Overall architecture of proposed research

3.2 Feature extraction model

The performance of classifier depends directly upon the choice of feature extraction approach employed on signals. Here the features are extracted from the obtained transformed signal which helps to predict the emotions with accurate manner. This process also helps to carry less memory consumption space and enhances the computational speed of the classifier. For

getting relevant and non-repetitive representation of wave files; temporal, spectral and energy based feature extraction is carried out in this phase. The working structure and mathematical description of those features are depicted below:

3.2.1 Spectral based feature extraction models

Spectral subset of features like Mel Frequency Cepstral Coeffients (MFCC), spectral roll-off, spectral flux, and spectral centroid are extracted for the transformed signal.

a. Mel Frequency Cepstral Coefficient

MFCC [26] is used as the fundamental music related feature because it discovers its applicability for modelling music. This frequency bands are logarithmically placed on a Mel-scale that measures the response of the human auditory system. It works better than linear spaced frequency bands. For generating the MFCC subset of features, the frequency scale signals are transformed to Mel-scale. Here the speech signal contains signals of tones with various frequency. The actual frequency f of each tone is measured in the frequency Hz and in the subjective pitch is Mel scale. The Mel frequency carried here is a linear frequency spacing less than 1000 Hz and a logarithmic spacing greater than 1000 Hz. 1 kHz tone is a pitch as reference point and the perceptual hearing threshold is 40dB above, and is defined as 1000 Mels. Therefore, to compute the Mels for a given frequency f in Hz we can use the Eq. (2).

$$Mel_{scale}(f) = 2595 * log\left(1 + \frac{f}{700}\right) \tag{2}$$

Here f denotes the frequency. After computing MFCC, some other spectral features are extracted. These features are computed by means of frequency spectrum of signal from time domain analysis and they possess single valued feature set.

b. Spectral Flux

This type of spectral frequency is otherwise known as Delta Spectrum Magnitude and it is the measure of change of power spectrum from each frames [26]. It is calculated by relating the two frames power spectrum. It is mathematically given in Eq. (3).

$$Spectl_{flux}(s) = \sum_{m=1}^{f} (N_s(m) - N_{s-1}(m))^2 \tag{3}$$

Here $N_s(m)$ and $N_{s-1}(m)$ represents the normalized magnitude at frame s and $s-1$ of Fourier transform and$m = 1tof$ is the frequency range.$Spectl_{flux}(s)$ represents the spectral flux.

c. Spectral Centroid

This spectral feature instructs the centre of gravity of the magnitude spectrum and it provides the brightness or clarity of sound. It is measured by the weighted mean of spectral frequencies. It is mathematically describes in Eq. (4):

$$Spectl_{centroid} = \frac{\sum_{m=1}^{f} N_s(m) * m}{\sum_{m=1}^{f} N_s(m)} \tag{4}$$

Here, $N_s(m)$ be the magnitude of transform at frame s at frequency$m$. If the obtained value for spectral centroid is lower means, more energy is located in lower frequency components. Similarly, it execute the vice versa operation [26].

d. Spectral roll-off

Spectral roll-off [26] is the frequency used to distinguish voiced from unvoiced music and is a measure of spectral shape. It is mathematically given in Eq. (5):

$$Spectl_{rolloff} = \sum_{m=1}^{f} N_s|m| \leq Q \sum_{m=1}^{M} N_s|m| \tag{5}$$

Where f represents the spectral roll off frequncy. Generally Q term attains 85% roll-off fraction.

3.2.2 Temporal based feature extraction approach

The temporal based features are extracted in the second feature extraction model. The essential feature extraction models used here are Zero crossing rate and compactness.

a. Zero crossing rate and compactness

Zero Crossing Rate (ZCR) is the basic noise indicator feature and is performed by counting the number of times frames the signal intersected the X-axis. It is defined as the count of time domain zeros is crossed. In other words it is described as the sign variation of signals per unit time. It is mathematically formulated as shown in Eq. (6):

$$ZCR = \frac{1}{2} \sum_{m=1}^{f} |sign(y(m)) - sign(y(m-1))| \tag{6}$$

Here $y(m)$ represents the signals frequency measure. The function *sign* returns 1 if $y(g)$ value is higher than 0. Likewise, this function obtains -1 value when $y(g)$ achieves value lease than 0. Similarly, $y(g)$ will be 0 if the *sign* function is zero. $y(m)$ represents time domain signal for frame s and M represents the length of frame of the speech signal. This feature helps to separate the voiced, unvoiced and silenced part of sound signal contained in the music. Because of its nature, this feature is broadly used in emotion-based music classification system. The next feature to extract in temporal feature is the compactness feature. This feature helps to find out the relatedness of elements in the audio frame. It is based upon the frequency domain feature set. It specifies the possible noises present in the signal so that it makes prediction process easier [6].

3.2.3 Energy related feature extraction

In the third stage of feature extraction process, energy based root mean square features are extracted and it is detailed below.

a. Root Mean Square

It is defined as the signals global energy and is evaluated by taking average root of the amplitude square. This measure is computes the error obtained per window basis. It is mathematically represented in Eq. (7):

$$RMS = \sqrt{\sum_{m=1}^{M} y_m^2} \tag{7}$$

Here the term y_m represents the signal measure at the point $\boldsymbol{m}$ and $\boldsymbol{M}$ is the frame length of speech signal. In the next phase, the extracted set of features are fed into the proposed classification algorithm to predict the different emotions related from music information [6].

3.3 Music based emotion classification framework

After extracting the music-related features, the next step followed is to train the proposed classification model to determine the relationship between the musical characteristics and the emotion label. In this research methodology, multi label emotions are recognized using proposed gradient descent based SNN classifier. The main purpose of adopting gradient descent to classification algorithm is that, it helps to optimize the error function occurs during the training process. For identifying the emotion with good accuracy rate, the obtained set of features are subjected as the input to SNN classifier.

3.3.1 Proposed gradient descent based Spiking Neural Network model

The extracted features are subjected to propose gradient descent based SNN structure for classifying multi-label emotions (nine emotions). Here gradient descent optimization technique is used to train the weight values of SNN classifier model [8]. In SNN model certain parameters are considered as significant for effective classification such as weight, bais and activation function. So, in order to determine the best weight value prior to processing gradient descent optimization is utilized. Using this optimization algorithm the best weight solution is obtained and given in SNN for achieving accurate classification with decreased error. The network employed in the spiking neurons denotes the unique computational approach which increases level of biological realism by using individual spikes. The spiking neuron present within the network creates a spike or action potential. This happens when a membrane state present internally crosses the threshold. The membrane that is present internally is termed as membrane potential. The neurons operation of spiking neural network is described as follows.

The j^{th} neuron get exceeded at the time interval u due to presence of spike train $G_k = \left\{u_k^{(1)}, u_k^{(2)}, , u_k^{(g)}\right\}$ which is arriving from its predecessorsΨ_j,then the presynaptic neuron which is represented as $k$belongs to the predecessors $k\epsilon\Psi_j \cdot$ here; $u_k^{(g)}$ represents the time interval during which the neuron $k$creates a spike. The input current denoted as J_j for j^{th} neuron is acquired by using Eq. 8.

$$J_j(u) = \sum_{k\epsilon\Psi_j} w_{jk} \sum_{g} \alpha\left(u - u_k^{(g)}\right) \tag{8}$$

Here w_{jk} represents the synaptic efficiency among the two neurons j and$k$. Similarly, α denotes the spike response function which is generated based on arrival of a single spike that charges

the membrane potential related to target neuron$j$. This role takes place in the post-synaptic function and can be given based on following Eq. (9).

$$\alpha(s) = \frac{r}{U_t} exp\left(-\frac{t}{U_t}\right)\Phi(t) \tag{9}$$

Here $\alpha(s)$ function is categorized based on amplification factor and this amplification factor is denoted as r and it declines exponentially with respect to constant timeU_t. The step function $\Phi(t)$ can be represented as

$$\Phi(t) = \begin{cases} 1, for\, t > 0 \\ 0, otherwise \end{cases} \tag{10}$$

The input neuron J_j goes beyond Φ threshold values, then j discharges and emits a new spike which creates the subsequent neurons. The diagrammatic representation of SNN adopted for emotion label classification is depicted in figure (Fig. 2).

The three layers namely input layer, hidden layer and output layer which are described as follows:

Layer 1- Input layer: This layer receives input music features corresponding to the different frame is converted to spikes using neural coding. The neurons contained in layer 1 is interlinked to a small number of neurons contained in layer 2.

Layer 2- Hidden layer: The hidden layer comprises of two neurons namely α andβ. Both receive input synaptic from input layer. It seems α and β represents excitatory and inhibitory type. These neuron types are connected to a larger number of neurons belonging to the same layer. From each neuron, half of the connections are associated with neuron α and other half connections are with β neurons.

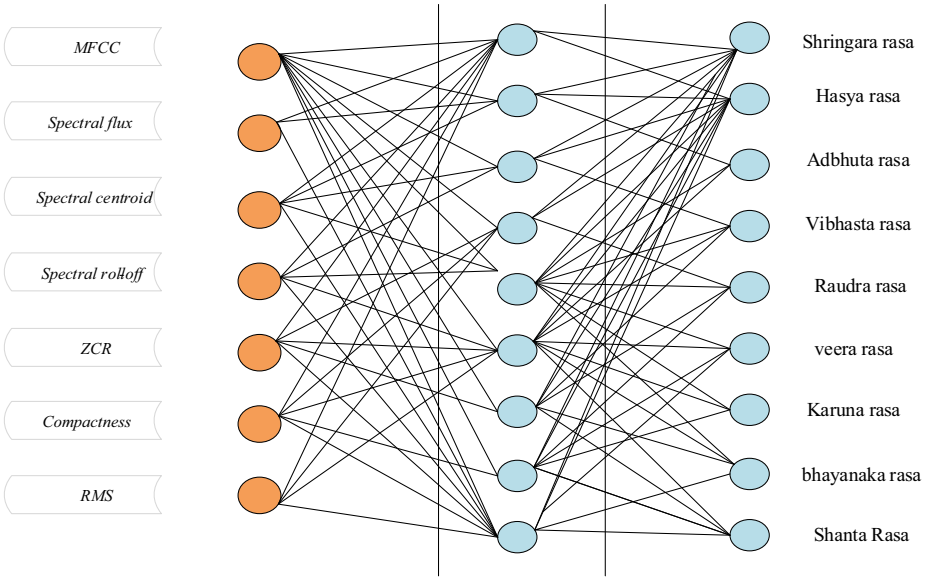

Fig. 2 Architecture of spiking neural network

Layer 3- Output layer: The neurons in the output layer receive connection from both α and β neurons from layer 2. These connections are defined at the time of training process. The output layer generates multi-label emotion with every possible rasa elicited from the segment of the raga.

3.3.2 Parameter optimization via gradient descent algorithm

In this optimization method, set of randomly generated weight values of proposed classifier algorithm are initialized. Subsequently, their performance are evaluated using the fitness function, which, in this case, is the accuracy value obtained when performing the activation with the training set. Consequently its error function also gets reduced while improving the accuracy. So that the traditional SNN model was altered by optimally choosing the weight values to reduce these overfitting issues. Thus the obtained best solutions by gradient descent was taken throughout the entire process. Gradient descent [13] is an optimization technique suitable for updating the parameters values in the training stage of neural network. It takes only the first derivative order for performing updation process in each iterations. The step size of each iteration to reach the local minimum depends upon the learning rate. The step by step procedure to be followed in gradient decent model is illustrated as follows:

Step 1: Initialization: The random weight values generated by SNN is initialized. The initialization is represented as follows:

$$E(u) = \{w_1, w_2, , w_k\} \tag{11}$$

Here $E(u)$ defines the initialization of random weight values. w defines the weight values generated in classification algorithm and k represents the number of terms considered for initialization.

Step 2: Fitness function: For obtaining best weight values for SNN, fitness function is determined and it is evaluated based on the below expression.

$$ff_{E(u)} = maximize(Accuracy) \tag{12}$$

$$Accuracy = \frac{\sum_{i=1}^{k} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + tn_i}}{k} \tag{13}$$

By doing this, best classification measure is obtained by maximizing the classification accuracy. At the same time, training error are reduced and the network becomes more effective for training the emotion classes.

Step 3: Updating: Based on the optimum weight values obtained from fitness evaluation, the training phase is carried out. The updating procedure of gradient descent model is depicted below:

$$w_{jk} = w_{jk} - \eta * \nabla . ff_{E(u)} \tag{14}$$

Here $ff_{E(u)}$ is the objective fitness function, $\in$ *weightparameters* and η determines learning rate which shows the step size to reach the entire iteration. During the updating process of each and every solution, the fitness calculation is evaluated for finding the most excellent solution among them.

Step 4: Termination criteria: Atlast, it satisfies the finest weight of SNN model by the gradient descent optimization approach. As a result of finding the optimal solution or best fitness function, the prediction model is qualified which means after training SNN structure, nine emotion labels was identified. Those nine labels are Shringara rasa, Hasya rasa, Adbhuta rasa, Vibhasta rasa, Raudra rasa, veera rasa, Karuna rasa, bhayanaka rasa and Shanta Rasa. The Shringara rasa which represents love and Hasya rasa which means laughter, belong to the first quadrant. Adbhuta rasa which means surprise and Vibhasta rasa which generates disgust, falls in fourth quadrant. Raudra rasa means anger and veera refers as heroism, falls in second quadrant. The melancholic shade of Shringara rasa is Karuna rasa and bhayanaka rasa represents fear, fall in third quadrant. The Shanta Rasa means the state of transcendence, calmness, serenity and it is placed at the center. Based on this principle, each layer is trained and the error get reduced.

The pseudo code of the proposed methodology is presented in the below pseudocode (Table 1),

3.3.3 Testing phase

Finally testing process is carried out for the remaining music dataset to evaluate the proposed methodology. Here, the classifier uses knowledge gained at the time of training phase to categorize the emotion label. The steps will be repeated until the whole dataset is scanned.

Table 1 Pseudo code

Pseudo code of proposed methodology
Input: $\mathtt{Dataset}\ d_n$ *which is pre-determined of* n *number of ragas*
Output: Emotion recognition model (Nine class);
1. *Perform pre-processing step by means of STFT using equation (1)*
2. *Extract the music related features from dataset* d_1
Based on spectral, temporal and energy based feature extraction techniques using equation (2-7)
*//*Create gradient descent SNN structure*//*
Optimizing weight values in SNN with gradient descent optimization algorithm using equation (11-14)
3. ***Train the dataset*** $\boldsymbol{d_1}$
4. *The wave format of the file in dataset* d_1 *is converted into corresponding spikes set* s_1
5. *Discover the representation of internal data from dataset* d_1
6. *Consider another dataset* d_2 *and its rasa has to be determined.*
7. ***Test the dataset*** $\boldsymbol{d_2}$
8. *Process the dataset* d_2 *using pre-processing state*
9. *Extract the features from the* d_2 *dataset*
10. *The wave format of the* d_2 *dataset is converted to corresponding spikes set* s_2
11. *Discover the representation of internal data from* d_2 *dataset*
12. *Compare the spikes* s_1 *with pre-determined spikes set* s_2
13. *Then, data prediction takes place*
14. *Evaluate the emotions or rasa recognition by pairing with the obtained trained values.*
15. *Stop the process*

4 Result and discussions

This section carries the experimental outcome of the proposed methodology. The implementation helps to show the superiority of proposed research. For classifying different human emotions; initially the input music audio clip is carried out to pre-processing phase. For reducing the lengthy audio clips into smaller frames, pre-processing steps are followed. Then those pre-processed output signals are subjected to the feature extraction stage where the most relevant music features are extracted. In the final stage, proposed classification algorithm classifies nine emotions and it was tested in the MATLAB tool. For analysing the performance of proposed research, several performance measures like Sensitivity, Specificity and Accuracy are evaluated. Additionally, to justify the effectiveness of proposed algorithm, it is compared with the existing techniques such as CNN (Convolutional Neural Network), DBN (Deep Belief network), SVR (Support Vector Regression), GMM (Gaussian Mixture Model), DBN-HMM (Deep Belief Network-Hidden Markov Model), SVM (Support Vector Machine), and DBN-SVM (Deep Belief Network- Support Vector Machine).

4.1 Dataset description

The dataset taken in the proposed methodology is in the .wav format collected from musical notes of the musicians. The dataset is available in https://zenodo.org/record/1257114#.XwUicygzbIV. This dataset is utilized for music emotion analysis.

4.2 Performance evaluation

For emotion recognition system from music notes, multi-label emotions are classified. The performance analysis of the proposed SNN and existing techniques like CNN, DBN, SVR, GMM, DBN-HMM, SVM and DBN-SVM are compared in this section to show the effectiveness of proposed methodology. The metrics used for performance evaluation are accuracy, sensitivity, specificity and error rate. The mathematical expression used for calculating these metrics in case of multiclass classification problem is illustrated below.

Accuracy

The accuracy is defined as the number of samples correctly detected as true positive or true negative from total number of samples. In case of multiclass it is calculated as average accuracy per class. The formula used for calculating accuracy for multi-class problem is given in Eq. (15)

$$Accuracy = \frac{\sum_{i=1}^{k} \frac{tp_i + tn_i}{tp_i + tn_i + fp_i + tn_i}}{k} \tag{15}$$

In Eq. (15) k represents the total number of classes, tp_i denotes truly positive, tn_i denotes truly negative, fp_i denotes false positive and fn_i denotes false negative.

Sensitivity

Sensitivity is otherwise known as recall. It is defined as number of samples correctly predicted as positive from total actual positive samples. The mathematical expression used for representing sensitivity for multi-class problem is given in Eq. (16)

$$sensitivity = \frac{\sum_{i=1}^{k} tp_i}{\sum_{i=1}^{k} (tp_i + fn_i)} \tag{16}$$

Specificity

Table 2 Overall comparison table for proposed and existing techniques with different measures

Techniques	Sensitivity	Specificity	Accuracy (%)
CNN	76.23	81.75	72.4
DBN	67.49	69.92	81.91
SVR	68.41	72.56	85.87
GMM	57.18	60.34	84.73
DBN-HMM	82.25	85.23	88.62
SVM	59.36	63.98	89.52
DBN-SVM	85.32	89.39	93.71
Proposed SNN	91.27	97.91	94.55

Specificity is termed as number of samples correctly identified as negative from total number of actual negative samples. The mathematical expression used for representing sensitivity for multi-class problem is given in Eq. (17)

$$specificity = \frac{\sum_{i=1}^{k} tn_i}{\sum_{i=1}^{k} (tn_i + fp_i)} \tag{17}$$

Error rate

Error rate is defined as average of per-classification error. And the formula used for calculating error rate is given in Eq. (18)

$$error\ rate = \frac{\sum_{i=1}^{k} \frac{fp_i + fn_i}{tp_i + tn_i + fp_i + tn_i}}{k} \tag{18}$$

The above mentioned metrics are estimated for both proposed as well as for existing method to perform comparison study for proving its effective functioning (Table 2).

By implementing spiking neural networks for emotion recognition or mood prediction from music as an input, its sensitivity, specificity and accuracy of recognition is analysed for multi-class problem through calculating the average for every metrics and it is determine to be better on comparison with other approaches. For instance, gradient based SNN model achieves 94.55% of accuracy in recognition of emotion which is better than other existing approaches. It is compared with existing techniques like Support Vector Machine (SVM), Convolutional

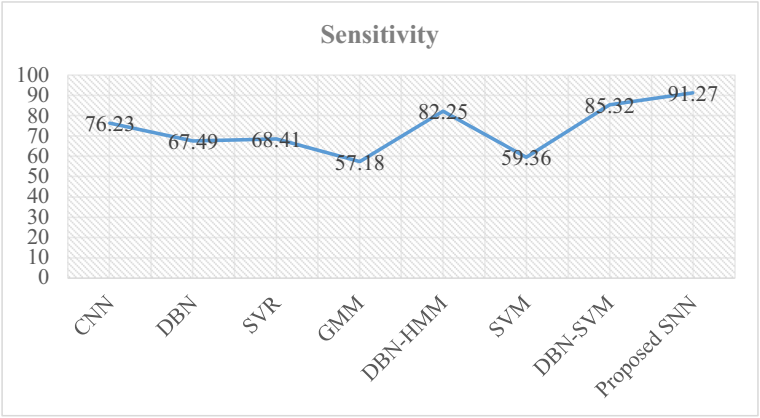

Fig. 3 Sensitivity acquired for proposed and existing techniques

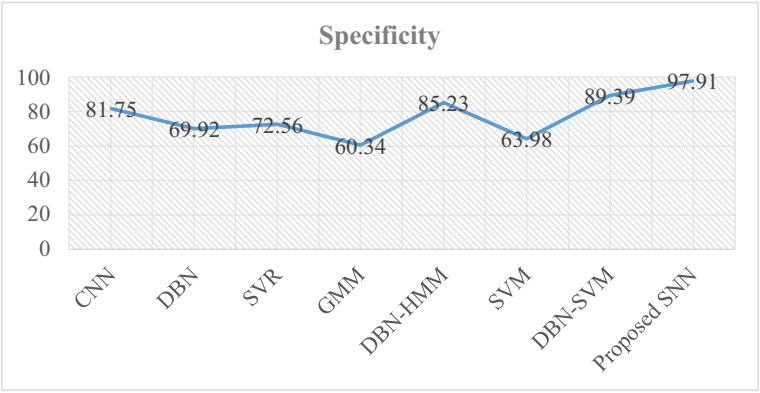

Fig. 4 Comparison plot for specificity measure

Neural Network (CNN), Deep Belief Neural (DBN) Network, Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), Support Vector Regression (SVR), and Spiking Neural Network (SNN).

The comparison plots are plotted in the upcoming figures for the clarity of proposed research work (Figs. 3, 4, and 5).

From the above comparison plot, the proposed SNN method is more reliable than any other existing technique. When equating sensitivity, specificity and accuracy, the proposed method demonstrates 94.55% accuracy and the existing techniques where CNN achieves 72.4%, DBN achieves 81.91, 85.87 for SVR, 84.73 for GMM, 88.62 for DBN-HMM, 89.52 for SVM, and DBN-SVM achieves 93.71 measures. Same like this, sensitivity and specificity measures are evaluated. It is clear that the proposed methodology attains better performance with higher accuracy when compared to another classifier algorithm.

Figure 6 shows the comparison graph of error rate achieved between the anticipated and traditional emotion-based classification techniques. From that, the error rate attained for proposed technique is 0.2, whereas it is higher for other conventional algorithms. Therefore, from the detailed analysis, it is clear that the proposed methodology attains better performance with less error rate when compared to other techniques.

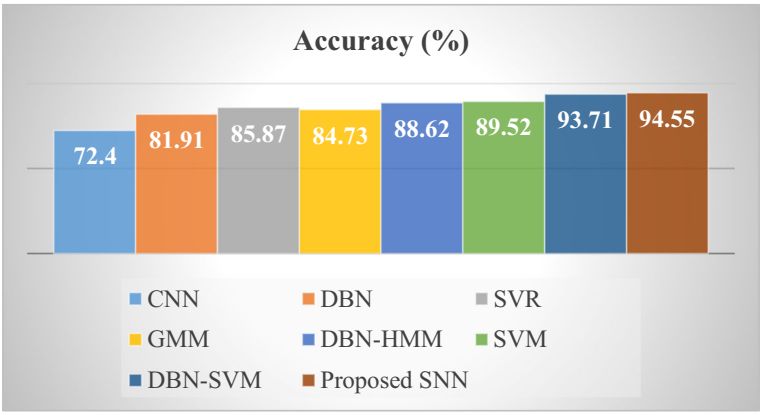

Fig. 5 Obtained accuracy for proposed and existing algorithms

Fig. 6 Error analysis for proposed and existing techniques

5 Conclusions

The main objective of the proposed methodology is to identify multi-label emotions from input music dataset. The collected dataset is initially subjected to pre-processing stage where the lengthy audio clips are segmented to smaller frames. Then feature extraction approach is performed for those segmented files whereas most relevant features are extracted. Afterwards, the extracted features are fed into the proposed gradient descent based SNN classifier. The proposed classification model helps to categorize multi-label emotions. The proposed methodology is compared with existing techniques like CNN, DBN, SVR, GMM, DBN-HMM, SVM and DBN-SVM. From the experimental outcome, it is observed that accuracy obtained by the proposed model is better than the existing algorithms. Our future works will be focused towards the development and tests of other texture features and other classification strategies for emotion classification.

Declarations

Conflict of interest There is no conflict of Interest between the authors regarding the manuscript preparation and submission.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informal consent Informed consent was obtained from all individual participants included in the study.

References

1. Antelis JM, Falcón LE (2020) Spiking neural networks applied to the classification of motor tasks in EEG signals. Neural Netw 122:130–143
2. Bajaj V, Taran S, Sengur A (2018) Emotion classification using flexible analytic wavelet transform for electroencephalogram signals. Health Inf Sci Syst 6(1):12
3. Baniya BK, Lee J (2017) Rough set-based approach for automatic emotion classification of music. J Inf Process Syst 13(2):400–416. https://doi.org/10.3745/JIPS.04.0032
4. Barthet M, Fazekas G, Sandler M (2012) Music emotion recognition: From content- to context-based models. Computer music modelling and retrieval, pp 228–252

5. Bashwiner DM, Wertz CJ, Flores RA, Jung RE (2016) Musical creativity "revealed" in brain structure: interplay between motor, default mode, and limbic networks. Sci Rep 6:204–282
6. Baume C (2013) Evaluation of acoustic features for music emotion recognition. In Audio Engineering Society Convention. Audio Engineering Society, 134
7. Bhatti AM, Majid M, Anwar SM, Khan B (2016) Human emotion recognition and analysis in response to audio music using brain signals. Comput Hum Behav 65:267–275
8. Buscicchio CA, Górecki P, Caponetti L (2006) Speech emotion recognition using spiking neural networks. In International Symposium on Methodologies for Intelligent Systems. Springer, Berlin, Heidelberg, pp 38-46
9. Capizzi G, Sciuto GL, Napoli C, Woźniak M, Gianluca Susi (2020) A spiking neural network-based long-term prediction system for biogas production. Neural Netw 129:271–279
10. Charles J, Lekamge LS (2019) The applicability of miremotion in emotion classification of Sri Lankan Folk Melodies. Available at SSRN 3496519
11. Chen A, Dai X (2010) Internal combustion engine vibration analysis with short-term Fourier-transform. In 2010 3rd International Congress on Image and Signal Processing, 9. IEEE, 4088-4091
12. Costa YMG, Oliveira LS, Koericb AL, Gouyon F (2011) Music genre recognition using spectrograms. IEEE International Conference on Systems, Signals and Image Processing(IWSSIP), pp 1– 4
13. Du SS, Zhai X, Poczos, Singh A (2018) Gradient descent provably optimizes over-parameterized neural networks. arXiv preprintarXiv:1810.02054
14. Eyben F, Salomao GL, Sundberg J, Scherer KR, Schuller B (2015) Emotion in the singing voice—a deeper look at acoustic features in the light of automatic classification. EURASIP J Audio Speech Music Process 2015(1):1–9
15. Furnkranz J, Hullermeier E, Mencia, Brinker K (2008) Multilabel classification viacalibrated label ranking. Mach Learn 73(2):133–153
16. Gyöngyössy NM, Domonkos M, Botzheim J, Korondi P (2019) Supervised learning with small training set for gesture recognition by spiking neural networks. In 2019 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, pp 2201-2206
17. Juslin PN, Laukka P (2010) Expression, perception, and induction of musical emotions: A review and aquestionnaire study of everyday listening. J New Music Res 33(3):216–237
18. Lin C, Liu M, Hsiung W, Jhang J (2016) Music emotion recognition based on two-level support vector classification. In 2016 International Conference on Machine Learning and Cybernetics (ICMLC) (1: 375-389). IEEE
19. Liu Y, Liu Y, Zhao Y, Hua KA (2015) What strikes the strings of your heart?—feature mining for music emotion analysis. IEEE Trans Eff Comput 6(3):247–260
20. Lokhande PS, Tiple B, Systems C (2017) Aframework for emotion identification in music: Deep learning approach. In: 2017International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, pp 262-266
21. Malheiro R, Panda R, Gomes P, Paiva RP (2016) Emotionally-relevant features for classification and regression of music lyrics. IEEE Trans Affect Comput 9(2):240–254
22. Meftah B, Lézoray O, Chaturvedi S, Khurshid AA, Abdelkader Benyettou (2013) Image processing with spiking neuron networks. Artificial Intelligence, Evolutionary Computing and Metaheuristics. Springer, Berlin, Heidelberg, pp 525–544
23. Misron MM, Rosli NB, Manaf NA, Halim HA (2014) Music motion classification (mec): Exploiting vocal and instrumental sound features. Recent Advances on Soft Computing and Data Mining, pp 539–549
24. Mo S, Niu J (2017) A novel method based on OMPGW method for feature extraction in automatic music mood classification. IEEE Trans Affect Comput 10(3):313–324. https://doi.org/10.1109/TAFFC.2017.2724515
25. Panda R, Malheiro RM, Paiva RP (2018) Novel audio features for music emotion recognition. IEEE Trans Affect Comput 11(4):614–626. https://doi.org/10.1109/TAFFC.2018.2820691
26. Pouyanfar S, Sameti H (2014) Music emotion recognition using two level classification. In 2014 Iranian Conference on Intelligent Systems (ICIS), IEEE, pp 1-6
27. Querlioz D, Bichler O, Dollfus P, Gamrat C (2013) Immunity to device variations in a spiking neural network with memristive nanodevices. IEEE Trans Nanotechnol 12(3):288–295
28. Rachman FH, Sarno R, Fatichah C (2018) Music emotion classification based on lyrics-audio using corpus based emotion. Int J Electr Comput Eng 8(3):2088–8708
29. Sanden C, Zhang JZ (2011) An empirical study of multi-label classifiers for music tag annotation. In: Proc. of the 12th International Society for Music Information Retrieval (ISMIR) Conference, pp 717–722
30. Schmidt EM, Turnbull D, Kim YE (2010) Feature selection for content-based, time-varying musical emotion regression in Proc. of the 11th ACMSIGMM International Conference on Multimedia Information Retrieval (MIR), 267–274

31. Tsoumakas G, Katakis I, Vla-havas I (2011) Random k-label sets for multilabel classifica- tion. IEEE Trans Knowl Data Eng 23(7):1079–1089
32. Wieczorkowska AA, Synak P, WRas Z (2006) Multi-label classification of emotions in music. In: Proc of Intelligent Information Processing and Web Mining, 35:307–315