An efficient supervised framework for music mood recognition using autoencoder-based optimised support vector regression model

Gaurav Agarwal¹ | Hari Om²

¹Research Scholar, Department of Computer Science & Engineering, IIT(ISM), Dhanbad, Jharkhand
²Assistant Professor, Department of Computer Science & Engineering, IIT(ISM), Dhanbad, Jharkhand

Correspondence
Gaurav Agarwal, Research Scholar, Department of Computer Science & Engineering, IIT(ISM)
Dhanbad, Jharkhand 826004, India.
Email: gaurav13shaurya@gmail.com

Funding information
National Natural Science Foundation of China (NSFC), Grant/Award Number: 51679176

Abstract
Music is the art of ‘language of emotions’. Recently, music mood recognition is an emerging task. An efficient supervised framework for music mood recognition using autoencoder-based optimised support vector regression (SVR) model is developed for the music emotion recognition. Our main intention is to increase the accuracy of emotion classification of music by considering text-dependent and non-text-dependent features. For the high level feature representation, stacked autoencoder is used with two hidden layers. Modified K-Medoid-based brain storm optimisation-based support vector regression (SVR_KMBSO) model is utilised for the emotion classification. Using the K-Medoid-based brain storm algorithm, the optimal parameters of the SVR are selected. The proposed framework utilises ISMIR2012 dataset and NJU_V1 dataset for English and for Hindi; online songs are also gathered and used for the music mood recognition. All the three datasets include songs based on four emotions like happy, angry, relax and sad. The experimental results are evaluated and compared with the existing classifiers including SVR, deep belief network (DBN) and Recurrent neural network (RNN). The proposed method SVR_KMBSO achieved high accuracy using three different datasets.

1 | INTRODUCTION

Music is one of the common ingredients to encourage our daily life. In the research community from the music information retrieval, music emotion recognition (MER) has attracted increasing attention. Recently with various perspectives of MER, significant research works are carried out on bimodal approaches, exploitation of lyrical information, automatic playlist generation, emotion variation detection and classification of song excerpts [1,2]. Almost every day, an enormous amount of music is published in this real-world to satisfy the music listener’s favourite styles and diverse interest. Recognising the emotion of the music clip is the intention of the MER [3]. Based on the emotion semantic, the MER is essential as it enables music retrieval, indexing and organisation [4]. For describing the music emotion, most of the MER systems adopt the valence-arousal [5].

MER model is divided into categorical models and parametric models in accordance with the psychological theory. By using the valence and arousal, the emotions in music are represented by the parametric approaches. Whereas, in categorical approach, the emotion labels or the adjectives are tagged with the songs such as peaceful, happy, calm, bored and angry [6,7]. MER highly concerns on developing the computational models to accurately predict the affective contents in each music signal. By employing machine learning methods, the MER model can be trained. It maps the perceived music emotion with the acoustic features [8]. Human expressions and emotions are not considered by most of the recommender systems. However, in our daily life, emotions have important influences.

To recognise internal cognitive processes and emotions of people, psychological signal method is more reliable than the facial expressions [9]. On the basis of lyrics, the classification of music emotion is called text classification. For text classification, there are various algorithms such as decision tree, k-nearest neighbour (KNN), Bayes, and SVM [10]. The music information retrieval sources are classified as audio, text and mixed audio and text. Triggering and communicating the listener’s emotions are identified as the major role of music.
Determining the emotional content from music in an automatic manner is considered as an essential task [11].

In several fields, the MER has many application values. For the automated music emotion classification and regression, various techniques have been proposed, and significant results have been achieved [12]. MER using computers has increasingly entered the perspective of people. Generally, MER is divided into three sections, which include emotion model, feature model and classification model. In MER, the selection of features of data is crucial. But for the researchers, it is challenging to get a strong idea about the process of MER because of the diversity and the complexity of the research systems. Hence, it is essential to compare and classify the techniques on the basis of text dependent and non-text dependent features utilised for the MER [13,14]. Some common issues arise when mapping the relationship between the emotions and music features.

Some of the issues include determining suitable feature extraction approach, and there is a correlation between the emotion and human preference, character etc. [15]. As in other recognition issues, the recognition techniques should be given along with the emotion model description. For the MER, various methods including multi-modal, ranking based and regression have been put forward [16,17]. In research community for the MER, a considerable amount of work has been devoted because of the human's response to music listening is mostly related to the emotion [18]. In our daily life, music is an important role, and during relaxation time, people enjoy music. The collection of songs increases every day in people's personal devices. There is a close relation between emotion and music so that people choose songs to listen, which suit their emotion [19,20].

The main contributions of this entire work are given as follows:

- The songs are converted into text and music without lyrics (karaoke). The features are extracted as both the text-dependent and non-text-dependent features. Also the results are evaluated and compared for songs, karaoke and lyrics by using three datasets by considering both English and Hindi.

- SVR_KMBSO-based method is introduced for the classification process in which the optimal parameters of the support vector regression (SVR) are selected by using the KMBSO algorithm. The selection of optimal parameters using modified optimisation algorithm is considered as the major contribution of our proposed approach, as it reduces the complexity of recognition process by performing clustering as initial phase during emotion classification. The clustering in BSO reduces the complexity and increases the speed of optimal parameter selection.

Remaining sections of this entire paper is organised as follows: Few related works are discussed in Section 2, Section 3 includes the problem definition, Section 4 provides the proposed methodology, Section 5 gives the experimental analysis and Section 6 contains the conclusion.

2 | RELATED WORKS

Some of the recent research works related to the music mood recognition are listed below.

In Dong et al. [21], a new technique named as bidirectional convolutional recurrent sparse network (BCRSN) was introduced for MER. With recurrent neural networks (RNNs) and convolutional NNs (CNNs), this recognition process was accomplished. The sequential-information-included affect-salient features (SII-ASF) was learnt by this method from the 2D spectrogram of audio signal. In this method, the three different processes such as ASF, feature extraction, and emotion prediction were combined for continuously predicting the emotions from the audio files. The weighted hybrid binary representation (WHBR) method was introduced in this approach for reducing the computational complexity. The regression-based prediction process was converted by this WHBR as a weighted combination for respective multiple binary classification issues. Music categorisation based on emotions with low level features and agglomerative clustering techniques was introduced by Sarkar et al. [22]. From the music signal, they considered only low level spectral features and time domain features. Based on the specific emotion, they carefully considered limited features instead of considering more features. K-means and Agglomerative clustering-based unsupervised method is considered for the process of classification. The method applied by Chaudhary et al. [23], for an automatic emotion recognition process is hash tag graph generation. Two different steps were included in this method: training and testing phases. Maximum recognition result was achieved by this method.

 Recommending the music based on emotions using wearable physiological sensors was developed by Ayata et al. [24], which recognises the user’s emotion from the music signal. From the multi-channel physiological signals, the recognition of emotion issue was observed as the prediction of valence and arousals. By the integration of physiological sensors and wearable computing device, the emotion is classified. The performance was evaluated on thirty two subjects with the fusion of features using KNN, SVM, random forest and decision tree algorithms. The outcomes showed that the proposed system’s accuracy can be integrated with any recommendation engine.

To accurately accomplish and perform mood classification for multimodal music, a new framework was developed by Hu et al. [25], which utilises complementary and multiple sources such as social tags related to music pieces, lyrics text and music audio. The experimental analysis showed that by using only the audio features, the proposed system outperformed with lyrics and audio. For the feature selection, the automatic methods proved to have the less feature space. The learning curves showed that when compared to the methods utilising either audio or lyrics, the hybridised method with the audio and lyrics features wanted shorter audio files and fewer training samples. Medina et al. [17] train the multi-layer perceptron (MLP) network using MediaEval database for emotion recognition from music. In this approach, initially, the pre-processing was applied, which balances and stratifies the dataset. It comprises three different classifiers for emotion recognition. The main contribution of our proposed approach, as it reduces the complexity of recognition process by performing clustering as initial phase during emotion classification. The clustering in BSO reduces the complexity and increases the speed of optimal parameter selection.

Remaining sections of this entire paper is organised as follows: Few related works are discussed in Section 2, Section 3 includes the problem definition, Section 4 provides the proposed methodology, Section 5 gives the experimental analysis and Section 6 contains the conclusion.
classification; they are random forest (RF), MLP, and SVM. Among these three, the MLP attained better recognition performance. Further, it integrates two binary classification techniques (binary classifier and one vs. rest [OvR]) for quadrant classification scheme. Finally, with the temporal annotation information, a dynamic classification having diverse time windows was executed.

Deep learning networks such as Support Vector Machine (SVM) and Random forest were utilised by Sarkar et al. [26] to perform the experiments on MER Traffic, Bi-Modal and Soundtracks dataset. An optimal results for the dataset was provided by the neural network, which combines the feature set for experimentation. Moreover, novel pre-processing technique was proposed in which it offers significant enhancement for the datasets. The VGGNet based on convolutional neural network was proposed to avoid the problems of feature design. In this work, a Valence-Arousal (V-A) model was utilised to automatically identify the emotion conveyed by the music. To train the classification models, the angle vectors of V-A plane was considered as a difficult task. So, a variant model of RNN was developed by Liu et al. [27] to extract the features from the melody. It was proved that the machine learning algorithm makes a good prediction of emotions from the music.

Gao et al. [28] proposed a hybrid neural network architecture (CRNN) for recognising the emotions from music, particularly for scratch-generated dataset. For such dataset creation, initially the main melody extraction approach was used. Then, the features from each music were extracted and provided as input to CNN. It learns the features and inputs the learnt features within RNN for emotion classification. The survey on existing techniques is shown in Table 1.

### Table 1: Literature review

| Author name and year | Methodology | Merits | Demerits | Research gap |
|----------------------|-------------|--------|----------|--------------|
| Medina et al. (2020) | Multi-layer perceptron | Attained low root mean square error (0.11) while predicting the valence and arousal. | It does not contains class balanced and quality dataset. Further the sound features and quality annotations are not found. | Further analysis by considering the variations in the length of time window is required. |
| Gao et al. (2020) | Recurrent neural network (RNN) | It effectively select the music having strong emotions. | It fails to determine the implicit feelings. | Emotion-wise classification is achieved only for four different emotions. |
| Sarkar et al. (2019) | Deep convolutional neural network | The network is lighten and modified by reducing different layers. | Attain minimum result for low arousal. | Time series nature corresponding to audio data is not considered. |
| Liu et al. (2019) | CRNN | Attained better results for homemade scratch generated music dataset. | Interference is high. | Fails to consider more appropriate and complete pre-processing approach. |
| Dong et al. (2019) | Weighted hybrid binary representation (WHBR) | Greatly diminish the training time and enhance the prediction accuracy. | Unavailability of proper dataset. | It fails to analyse the text based features. |
| Sarkar et al. (2019) | K-means and agglomerative clustering | Wide range of features are obtained. | Lack of optimal training data. | The general observation is that it is quite difficult to capture and map the emotion in terms of low level features. |
| Choudhary et al. (2019) | Hash tag graph (HTG) generation method | Efficiently predict the emotions. | Not possible to categorise a music clip to a single emotional class. | Deep learning architecture may be deployed to learn the features. |
| Ayata et al. (2018) | Emotion recognition with wearable computing devices | Enhance the accuracy of the resulting music recommendations | Failure in one wearable sensor creates the performance degradation. | It requires multiple sensors for performance enhancement. |
| Hu et al. (2017) | Multimodal music mood classification framework. | Examines information source on classifying music pieces across mood categories. | Limited the generalisability of the findings to a variety of music. | The classifier allows the solutions to fall in local optimum. |

Abbreviation: SAE, stacked auto encoder.
other research works. Therefore, to improve the recognition accuracy in this method, a novel optimisation-based SVR is introduced. In this method, along with English songs, the Hindi songs are also considered for analysis.

Music is a digital resource, and now it is extensively available in various media some of them are pen-drives, mobile phones, websites, digital audio players etc. The volume available for music makes it difficult for the listeners to choose a desired piece of music. Therefore, it is essential to develop an effective technique to facilitate an efficient searching for desired music. By taking all this in consideration, here a novel optimisation-based SVR is introduced to attain a high accuracy while recognising the music mood.

4 | PROPOSED METHODOLOGY

Normally, music is associated with the unique content of emotions, which is often referred as music mood. State-of-the-art methods concentrate on music piece content analysis, which considers the music mood as important metadata. The main goal of proposed work is to construct a music emotion classification framework for both Hindi and English music. For robust feature extraction, the key features such as intensity, rhythm, and timbre are considered for mood classification task. Further, the emotional expression-associated acoustic features are extracted through Mel-frequency cepstral coefficients (MFCC), octave-based spectral contrast (OSC), statistical spectrum descriptors (SSDs), Chromagram, Daubechies Wavelet Coefficient Histograms (DWCH). Also, the text dependent features are extracted by N-grams, Term frequency-inverse document frequency (TF-IDF) and total emotion count. The automatic supervised music mood classification framework is developed with deep learning-based network model. Initially, the high-level of features from the input data is extracted by stacked auto encoder (SAE) model. With these extracted features, the target emotions are accurately predicted by SVR-based recognition technique. Furthermore, the KMBSO algorithm is introduced along with SVR to improve its performance. The optimal parameter selection for SVR is effectively performed by this KMBSO algorithm. The schematic structure for this proposed recognition technique is shown in Figure 1.

Compared to the other existing works, the advantage of proposed work is after extracting the features from music signal based on text-dependent and non-text-dependent manner, and they are fed into the stacked autoencoder to obtain higher level feature representation to avoid emotion class level confusion. Further, the higher level feature representation is trained via SVR model. The SVR model hyper parameters are selected through efficient KMBSO algorithm, which leads to enhance the recognition performance of music mood classification.

4.1 | Feature extraction

4.1.1 | Text-dependent features

In songs, many types of textual features can be mined from the lyrics, which are very rich resources. Some of the most common features used for the classification of text are considered, which comprises N-gram, TF-IDF and total emotion count. The N-gram, TEC and TF-IDF are used for extracting the emotions from words.

a. N-grams

In various classification tasks, comprising classification of emotion and sentiment, the N-grams are the most standard level features. It is simple and highly scalable. We used unigram and bigram to construct the feature vector for the music emotion classification.

\[
P(W_n | W_1 W_2 W_3 ... W_{n-1}) = \frac{c(W_1, W_2, ..., W_n)}{c(W_1, ..., W_{n-1})} \tag{1}
\]

![Figure 1: Flowchart for proposed methodology](image-url)
where \(c(W_n)\) indicates the occurrence count of word \(W_n\).

b. Term frequency-inverse document frequency

It is normally represented as bag-of-word scheme, as the collection of words from the entire document is included in this TF-IDF [29]. TF-IDF is an efficient and simple technique. This feature basically assumes that the word that is highly significant is repeated number of times in the whole document; meanwhile, such word is rarely applied in other documents. In this, the initial assumption is related with TF feature, and the latter assumptions are related with IDF feature. The number of times that the word \(i\) appear in the document \(j\) is represented by parameter, \(tf_{ij}\). If the value of parameter, \(tf_{ij}\) is found high, then that particular word is found large number of times in document. The word \(i\) that is found number of times in the document \(j\) contains high TF and small IDF value.

\[
TF - IDF = tf_{ij} \times idf_i
\]

Where

\[
tf_{ij} = \frac{\text{frequency of term } i \text{ in document } j}{\text{total words in document } j}
\]

\[
idf_i = \log \left[ \frac{\text{total count of documents}}{\text{count of documents with term } i} \right]
\]

c. Total emotion count

The number of words that are associated with an emotion in the entire document is captured by this feature. Only a popular emotion context of a word that are recommended by the lexicon is captured by this feature (i.e. the emotion that having maximum score in the lexicon). But, these words not only show single expression is provided by all but also they show multiple expressions, for example, the word 'beautiful' which is normally applied in both joy and love emotion. Therefore, for such word, both joy and love emotions are count as one, and other emotions are count as 0 (based on the lexicon Lex scores). To avoid this issue, the feature that includes the relation of both word and multiple emotions needs to be developed. The features obtained for \(j\)th emotion is obtained using Equation (5),

\[
d_{TEC}(e_j) = \sum_{w \in d} I(e_j = \arg \max \text{Lex}(w, k)) \times \text{count}(w, d)
\]

where \(I(\cdot)\) represents the indicator function and \(e_j\) represents the feature value for \(j\)th emotion. TEC only collects the word those having highest score in lexicon [30]. The number of occurrences of word \(w\) in document \(d\) is represented as count \((w,d)\).

4.1.2 Non-text dependent features

The attributes of music are normally categorised into four to eight different categories; each one represents different concepts. The eight different categories are tone colour (associated with timbre), dynamics, harmony, rhythm, musical texture, expressive techniques, musical form, and melody. Two different understandings are provided by these categories; they are (i) where do the features that related to particular emotion belong (ii) and which category contains less computational models while performing the musical feature extraction from particular emotion. In this method, eight different non-text-dependent features are employed; they are rhythm, timbre, intensity, chromagram, MFCC, OSC, SSDs, and DWCH.

a. Rhythm

Music contains sequence of notes that may get changed over different time; however each contains its own specific duration. Therefore, determining the note duration statistics is identified as an obvious computational metric. The three different rhythm features that are closely related with the response of human emotions are rhythm regularity, tempo, and rhythm strength. Meanwhile, three different and possible categories are considered in this method to capture the dynamics related to duration and its changes; they are long, medium, and short notes. But the existing techniques determine the mean and standard deviation for the duration of all these notes to define these three ranges. The qualitative duration for note, is represented as \(ND_i\).

\[
\mu = \text{mean}(nd_i) \quad (6)
\]

\[
\sigma = \text{std}(nd_i) \quad (7)
\]

\[
ND_i = \begin{cases} 
\text{long}, & \text{if } nd_i \leq \mu - 0.5\sigma \\
\text{medium}, & \mu - 0.5\sigma < nd_i < \mu + 0.5\sigma \\
\text{short}, & nd_i \geq \mu + 0.5\sigma 
\end{cases} \quad (8)
\]

b. Timbre

Timbre is also referred as tone or colour of sounds. One of the basic element in music is timbre, which determines the quality of sound. It is nothing but the set of subjective opinion developed by individuals while looking onto the sound that is found independent from amplitude (loudness) or the frequency (pitch). Only frequency and amplitude are identified as the dominant attributes for timbre; due to this limitations, the timbre definition is viewed as a most subjective and varying one.

\[
T = Q_l - Q_i \quad (9)
\]
where $Q_i$ and $Q_j$ are the qualities of two sounds and $T$ denotes the timbre.

c. Intensity

Most essential feature for emotion recognition is intensity, as it is highly correlated by arousal. Therefore, it is widely applied for arousal dimension classification. The amplitude measurement accurately determines the intensity level and few descriptive features such as minimum energy rate; loudness centroid is included in song intensity. These descriptive features are extracted in frame due to the presence of short-time dynamic behaviour in intensity. The summed signal spectrum and the spectrum distribution from several sub-band are composed within the intensity feature of several frame, which is defined as in Equation (10),

$$ I(n) = \sum_{k=0}^{\infty} A(n,l) $$

where $I(n)$ is the intensity of the $n^{th}$ frame and $A(n,l)$ is the absolute value of the $l^{th}$ FFT coefficients of the $n^{th}$ frame.

d. Mel-frequency cepstral coefficients

MFCCs are originally designed for speech recognition, and they have been shown as the most informative feature domains for MER. This acoustic feature is most extensively applied in various audio and speech processing. MFCCs basically represent the low-dimensional warped spectrum based on mel-scale. This process is accomplished in frame-by-frame manner, after that the each frame is converted into a feature vector of single N-dimensional [31]. The number of samples in each frame is greater than the value taken as $N$. It provides the data required to process by the back end system in which it minimises the quantity of data. In feature extraction, the audio input is converted into a vector sequence $X = [x_1, x_2, ..., x_N]$, where $k$ indicates the frame index and $x_k$ denotes the N-dimensional vector.

Using the following steps, the MFCCs are determined

(i) First one is pre-emphasising the speech signal.

(ii) The speech signal is separated into a large number of frames with size 20 ms and shift 10 ms, then the hamming window is applied over these each frames.

$$ P_k(a) = \frac{1}{n} |DFT_k(a)|^2 \quad 1 \leq a \leq A $$

Here, $A$ represents the discrete fourier transform (DFT) length, and the power spectrum of $k^{th}$ hamming window is represented as $P_k(a)$. $n$ represents the sample number.

(iii) By applying DFT, the magnetic spectrum is computed for each windowed frame. The DFT of $k^{th}$ hamming window is estimated as:

$$ DFT_k(a) = \sum_{x=0}^{n-1} s_k(x)e^{-j2\pi \alpha x} $$

where the time domain signal of the $k^{th}$ hamming window is indicated as $s_k(x)$.

(iv) Compute the Mel spectrum obtained by passing the DFT signal through Mel filter bank.

(v) To the log Mel frequency coefficients (log Mel spectrum), DCT is applied to obtain the required MFCCs.

In this method, the dynamic and static features of music signal based on the 39th order MFCC is captured. Moreover, 39 dimensional MFCC feature vector of each frame is obtained with 13th order acceleration (delta-delta) coefficients, 13th order static coefficients and 13th order delta coefficients.

e. ctave-based spectral contrast

OSC feature is found in 14 dimensions, and it applies 7 octave-based bands [32]. The spectral valley, spectral peak, and their difference in each sub-band are considered by this OSC feature. The difference between spectral peak and spectral valley provides the contrast. The values that are found in each sub-band is sorted in ascending order by spectral valley, and finally, it sums the initial 2% of bandwidth value. Spectral peak also performs the similar sorting process but in descending order [31]. For extracting the OSC features, initially apply the Fast Fourier Transform (FFT) on each audio frame to generate the spectrum. Then, subdivide such spectrum into sub-bands with a set of octave scale filters. $P_{a,1}, P_{a,2}, ..., P_{a,Na}$ is the power spectrum in the $a^{th}$ sub-band, and $Na$ represents the number of FFT frequency bins in $a^{th}$ sub-band. The spectral valley and spectral peak of $a^{th}$ sub-band is given in Equation (13)

$$ \text{Peak}(a) = \log \left( \frac{1}{aNa} \sum_{i=1}^{Na} P_{a,i} \right) $$

$$ \text{Valley}(a) = \log \left( \frac{1}{aNa} \sum_{i=1}^{Na} P_{a,Na-i+1} \right) $$

where $a$ represents the neighbourhood factor which is set as 0.2. The difference between spectral peak and spectral valley provides the spectral contrast (i.e. $SC(a) = \text{peak}(a) - \text{valley}(a)$). The feature vector contains the spectral vector and spectral contrast of all sub-bands. Therefore, the OSC of feature vector is represented as

$$ X_{osc} = [\text{Valley}(0), ..., \text{Valley}(B-1), SC(0), ..., SC(B-1)]^T $$

where $B$ represents the number of octave scale filters.
TABLE 2 Algorithm for SAE training

| Algorithm: training of SAE |
|--------------------------------|
| Step 1: Initialise bias = 0, and weight matrix randomly between, \([-r, r]\), where \(r\) is the constant whose value may depend on the total neurons in the hidden layers. |
| Step 2: Set, \(\Delta W_l = 0\), and \(\Delta b_l = 0\) for all \(l\). |
| Step 3: Then, apply back propagation algorithm to compute weight and bias parameters, \(\Delta W_l := \Delta W_l + \delta_{l+1}(a_l)^T\) \(\Delta b_l := \Delta b_l + \delta_{l+1}\). |
| Step 4: Update weight and bias \(w_i = w_i - \varepsilon \left[ \frac{1}{m} \Delta (w_i) + \lambda (w_i) \right] \) \(b_i = b_i - \varepsilon \left[ \frac{1}{m} \Delta (b_i) \right] \) where \(\varepsilon\) represents the learning rate. |

f. statistical spectrum descriptor

During music and audio processing, the SSD is frequently related with timbral texture. The hanning window is applied for each spectral shape function to divide the data into short-overlapping segments and also to evaluate the magnitude DFT. It is a 4D feature, which contains spectral centroid, spectral flux, spectral rolloff, and spectral flatness in order.

\[ S_c = \frac{\sum_{k=1}^{k} f_k S_{1k}}{\sum_{k=1}^{k} S_{1k}} \]  \( (15) \)

\[ S_{fx} = \| S_i - S_{i-1} \|^2 \]  \( (16) \)

\[ S_r = 0.85 \sum_{k=1}^{k} S_{1k}^2 \]  \( (17) \)

\[ S_f = \frac{\left( \prod_{i=1}^{n} P_k \right)^{\frac{1}{2}}}{\frac{1}{n} \sum_{k=1}^{n} P_k} \]  \( (18) \)

where the centroid is denoted as \(S_c\), which identifies the weighted average of the spectrum; the flux is denoted as \(S_f\), which identifies the Euclidean distance among successive spectral frames; the roll-off is denoted as \(S_r\), it provides frequency, which is found less than the total spectral energy; and the flatness measure is denoted as \(S_f\), which quantifies the closest relationship between the uniformly distributed spectrums. \(f_k\) is the frequency of bin \(k\), \(S_{1k}\) is the strength of bin \(k\) in frame \(i\), \(P_k\) denotes the magnitude of bin \(k\), and \(n\) is the total number of bins in a band.

g. Chromagram

A spectrogram of 12 frequency bins, which represents the 12 semitones, is named as Chromagram or Choma. Chromagram is a well-defined method that is widely applied to determine the pitch class component in less duration. It is nothing but a circular form of the spectrogram that are logarithmically warped. Within this version, the chroma frequencies of various octaves are gathered and summed for evaluating the energy at the each 12 pitch classes. Sometimes, the indication of the entire modality and musical keys are attained by this features.

\[ S(t, c) = \sum_k \left( t^2 c^2 + k \right) \]  \( (19) \)

where \(S(t, c)\) indicates the time-chroma distribution, \(c \in [0, 1]\) and \(k \in Z\), in which \(Z\) indicates integers.

h. Daubeichies wavelet coefficient histograms

The wavelet coefficients of the Daubechies is used since they reflect the local regularity present in the signal being analysed. Daubechies wavelet filter having seven decomposition levels is applied in this method. After decomposition process, the histogram for the wavelet coefficient at each subband is developed [33]. The waveform variations are perfectly approximated at several subband by the coefficient histogram. Its moments then uniquely characterise the probability distribution from the probability theory. Next, the initial three moments from the histogram are applied to describe the waveform distribution. Three different moments that are applied for this characterisation are the skewness, the average, and the variance of each subband. DWCH can be computed by using the scaling function.

\[ \varphi(x) = \sum_{k=-\infty}^{\infty} a_k \varphi(2x - k) \]  \( (20) \)

where \(\varphi(x)\) is the scaling function and \(a_k\) is the filter coefficient.

FIGURE 2 Stacked autoencoder with two hidden layers
4.2 | Deep Neural Network-based emotion classification

4.2.1 | Stacked autoencoder-based high level feature representation

The text-dependent and text-independent features are extracted, and then that extracted features are provided to SAE for extracting high level features because the presence of low level features may reduce the performance of classification process. To avoid that, the SAE is introduced, which extracts the high level features and provide that extracted features to SVR for emotion recognition. AE is a type of unsupervised learning approach, which contains three different layers: input, hidden and output layers. The output attained from each layer is given as an input to the successive layers [34]. The hypothesis function \( \tilde{h}_{w,b}(x) = x \) (where \( w \) is weight and \( b \) is bias) having unlabelled data is learnt by SAE. The cost function for SAE is defined in Equation (21),

\[
J(w, b) = \frac{1}{m} \sum_{i=1}^{N} \frac{1}{2} \| h_{w,b}(x_i) - \hat{x}_i \|^2 + \frac{\lambda}{2} \sum_{j=1}^{n} \sum_{j=1}^{m} w_{ji}^2 + \beta \sum_{j=1}^{n} KL(\rho_\|\tilde{\rho}_j) \tag{21}
\]

\( KL(\cdot) \) represents the Kullback-Leibler divergence function, \( \beta \) represents sparse penalty term, and \( \rho \) is sparsity parameter. The average activation function subject to hidden units is represented as \( \tilde{\rho}_j \). The back propagation (BP) with batch gradient descent optimisation is introduced in SAE to compute the gradients of each learning weights (Table 2). Then, for each node compute the error \( \delta_l^{(i)} \) (where \( l \) represents layer and \( i \) represents each node).

Step 1 SAE determines \( w \) and \( b \) in the input layer (which contains extracted features).

Step 2 Then determine the feature size \( (a) \) from the first hidden layer by applying \( z = w \cdot data + b \) and \( f(z) \) where \( f \) represents the sigmoidal function.

Step 3 Then rank the features \( (a) \) with labels and drop the most least features.

Step 4 Repeat Steps 2 and 3 in \( 2^{nd} \) hidden layer by giving \( a' \) as input and obtain \( b \) and \( b' \).

Step 5 Also repeat steps 2 and 3 for \( 3^{rd} \) hidden layer by giving \( b \) and \( b' \) to obtain \( c \) and \( c' \).

The probability of output values are statistically estimated by SVM classifier. This classifier uses the features that are learnt by last hidden layer to learn the all bias and weight parameters. The architecture of deep neural network (DNN) has two hidden layer, one input layer and one output layer are shown in Figure 2.

After learning the bias and weight parameter in the output layer, the fine-tuning for all \( w \) and \( b \) parameters is performed simultaneously in the entire network using KMBSO algorithm. SAE is considered as single model, but it has various layers, and the fine-tuning process in all these layers use back propagation algorithm to enhance the weight parameter. Based on labelled training samples, the bias and weight parameter of entire network are learnt by this standard back-propagation algorithm, which mainly aims in reducing the misclassification errors. The high level features that are learnt using the SAE is given as input to the SVM (output layer) for emotion classification. The data from various classes are separated by SVM using a hyperplane. The algorithm for DNN training is shown in Table 3.

4.2.2 | Emotion classification using SVR_KMBSO

The SVR maps applied nonlinear mapping to map the input data within the high-dimensional feature space. The linear regression problem obtained in this feature space is solved by this mapping process. Based on the available dataset, the regression approximation determines the function. The main aim of this modelling process is to determine a regression function that accurately performs the output prediction for another (new) set of input-output samples. Figure 3 illustrates the fundamental idea behind the SVR.

Identifying the smooth regression function \( f(x) \) is considered as the main objective of this process. At the same time, the maximum deviation that attained from all the training
targets needs to be maintained below $\varepsilon$. The subsequent problem is solved to obtain the SVR model

$$\min \|w\|^2 + C \sum_{i=1}^{l} (\xi_i + \xi_i^*)$$

$$\begin{align*}
\{ & y_i - \langle w, x_i \rangle - B \leq \varepsilon + \xi_i, \\
& \langle w, x_i \rangle + B - y_i \leq \varepsilon + \xi_i^* \\
& \xi_i, \xi_i^* \geq 0
\end{align*}$$

(22)

The nonlinear decision hyper-surface is constructed over the SVR input space by kernel function. Normally, the high result on prediction performance is attained by Gaussian function. Based on nonlinear kernel function, the two problems are stated in Equations (22) and (23)

$$\text{maximize } L(\alpha) = -\frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j^* \langle k(x_i, x_j) \rangle$$

$$\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0, \quad \alpha_i, \alpha_i^* \in [0, C]$$

(23)

where the Lagrange multipliers are represented as $\alpha_i$ and $\alpha_i^*$, and $k(x_i, x_j)$ represents the kernel function. In the above equation, $(\alpha_i - \alpha_i^*)$ is non-zero if $x_i$ is a support vector. In this method, the Gaussian function is taken as the kernel function. The Gaussian kernel is defined as in Equation (24),

$$k(x_i, x) = \exp\left(-\gamma \|x_i - x\|^2\right)$$

(24)

where $\gamma$ is a positive constant.

The parameters that are defined by user are $C, \varepsilon$ and $\gamma$, and based on these parameters, the accuracy of entire system gets altered. The SVR parameters need to set carefully; therefore, this SVR model can perform more effectively. Therefore, parameter selection is consider as a significant step in SVR model. Then, with the selected parameters initially, train the SVR model. After completing training process, test the SVR model with valid parameters to obtain the accurate solution. This process is found time consuming and highly based on luck.

Therefore to minimise the consumption of time and for efficient parameters selection, K-Medoid based brain storm optimisation (KMBSO) algorithm is proposed. In this algorithm, there are two steps, which include clustering and new individual generation. For clustering the individuals, k-medoid approach is used in this proposed method. The proposed SVR_KMBSO model dynamically optimises the parameter values of SVR through the KMBSO algorithm.

**Hyperparameter optimisation using K-Medoid based brain storm optimisation**

Initially, the parameter values are generated randomly in the range of zero to infinity. These parameter values are clustered by using K-Medoid clustering. The ‘k’ information things are selected as beginning medoids by this clustering process to represent the $k$ clusters. Large number of residual
things are included within this cluster, which has its medoid nearest to them. Another medoid from that point is resolved to perfectly represent the cluster. Then, the K clusters that are mainly focussed on medoids are developed. In this K-Medoid clustering, the distance between the medoid and each individual is calculated by using Equation (25).

TABLE 6 Performance comparison of proposed method for various datasets

| Performance Metrics | ISMIR2012 | NJU_V1 | Hindi |
|---------------------|-----------|--------|-------|
|                     | Song      | Karaoke| Lyrics| Song | Karaoke | Lyrics |
| Accuracy            | 98.79     | 97.40  | 97.23 | 97.88 | 96.82   | 95.76  |
| Precision           | 98.77     | 97.41  | 97.22 | 97.87 | 96.64   | 95.66  |
| Sensitivity         | 98.78     | 97.38  | 97.21 | 97.93 | 96.74   | 95.75  |
| F1-score            | 98.77     | 97.4   | 97.25 | 97.89 | 96.69   | 95.70  |
| Specificity         | 99.60     | 99.13  | 99.08 | 99.29 | 98.95   | 98.58  |
| FPR                 | 0.40      | 0.87   | 9.23  | 0.71  | 1.05    | 1.45   |
| Error               | 1.21      | 2.60   | 2.77  | 2.12  | 3.18    | 4.24   |

FIGURE 4 Confusion matrices for ISMIR2012 dataset. (a) Song, (b) Karaoke, (c) Lyrics
where $D$ denotes the distance between the medoids and the data points, $m_i$ denotes the medoids and $d_i$ denotes the data points.

In this proposed method, the number of clusters is 3 (\(K = 3\)) since we have to optimise three parameters. One cluster is for insensitive loss function, one is for the selected Gaussian kernel function and the other is for the regularisation parameter. After clustering the data points, random values are generated for each data point in each cluster in the range of 0 to 1. Based on the priority, from each cluster, a data point is selected for each iteration and performs the SVR classification. The fitness function for this optimisation is given in Equation (26), and the objective function is given in Equation (27).

**Fitness definition**

The k-fold cross-validation (CV) applied on the training samples is identified as an effective approach in machine learning. This ensures that this method accurately predict the training dataset, meanwhile it analyse whether any of the selected parameters results over-fitting. The value of $k$ is practically set as 5–10, which maintains the trade-off of prediction accuracy and computational cost. In this method, the five-fold CV error referred as mean absolute percentage error (MAPE) is introduced in SVR to measure the training error, which is determined using Equation (25):

$$\text{MAPE_{cross validation}} = \frac{\sum_{i=1}^{l} |a_i - p_i|}{l} \times 100\%$$

where $l$ represents the number of training data samples; and the actual and predicted value is represented as $a_i$ and $p_i$, respectively. The solution those having less MAPE_{cross validation} 

**FIGURE 5** Confusion matrices for NJU_V1 dataset. (a) Song (b) Karaoke (c) Lyrics

\[
D = \sum_{i=1}^{n} (m_i - d_i)^2
\]

(25)
contains smaller fitness value; due to this, the survival rate of such solutions may get reduced in successive generations.

After processing all the individuals in the clusters, new data points are generated for every individual data point by using the below Equations (28) and (29)

\[
d_{i}^{\text{new}} = d_{i}^{\text{old}} + \delta(T) \times R, \quad i = 1, 2, \ldots, d_n
\]

\[
\delta(T) = \log \text{sig} \left( \frac{0.5 \times I - c}{C} \right) \times R
\]

where \(d_{i}^{\text{new}}\) indicates the new data points, \(d_{i}^{\text{old}}\) indicates the old data points, \(\delta(T)\) indicates the function of step size, \(R\) indicates the generated random value for the old data point, \(d_n\) indicates the total number of data points within the cluster, \(I\) indicates the total iteration number, \(c\) indicates the current iteration number and \(C\) indicates the coefficient for changing the step size function's slope.

The random values are generated in the range of 0 to 1; then, the ranking process is performed, and finally, the classification process is carried out by using the SVR classifier, and the fitness values are evaluated and stored in the array. Same process will be repeated till reaching maximum iteration. Finally from the stored values, a best parameter set is selected, which provides the minimum fitness value. Using the optimal parameters, the emotion classification is performed.

5 | EXPERIMENTAL ANALYSIS

This proposed mood recognition process is implemented in MATLAB 2017a platform, and the performance such as accuracy, precision, sensitivity, F1-score, false positive rate (FPR), specificity and error are evaluated with respect to text-dependent and non-text-dependent manner and compared with other classifier models such as SVR, recurrent NN (RNN), and

![Confusion matrices for Hindi dataset.](image)

**Figure 6** Confusion matrices for Hindi dataset. (a) Song (b) Karaoke (c) Lyrics
In this work, we have used one dimensional features. The parameter settings of proposed algorithms are shown in Table 4.

5.1 Dataset description

The proposed method is implemented in three datasets in which two datasets for English and one is for Hindi. The description for the three datasets are given below.

**ISMIR2012:** The proposed framework utilises Popular Music Emotion Dataset — ISMIR2012 dataset for English (western song), and for Hindi, online songs are collected and applied in this method for the classification of music mood. In the English dataset, totally there are 2886 songs in which 746 are happy songs, 759 are sad songs, 636 are angry songs and 745 are relax songs. In our Hindi dataset, there are 1037 songs in which 200 are happy songs, 216 are sad songs, 283 are angry songs and 338 are relax songs.

**NJU V1:** 777 music clips are included in this dataset each depicts four different moods: happy, relaxed, sad, and angry. From that 777 music clips, 400 clips are taken for training, and 377 clips are taken for testing process. These clip samples are distributed in four different mood categories. The music clips present in this dataset are provided with a plain text (.txt) file, which contains the lyrics of each music clips along with the time tags (i.e. the time offset of the sentence relative to the start of the music). In each training/testing sample set, four information fields are separated by colon. It includes music number in the package, music title, performer’s name and duration of music.

**Hindi dataset:** For Hindi, there is no dataset available with these four emotions. So, we have collected the songs from online sources for these emotions and used as Hindi dataset.
The song count in each emotion for the three datasets is provided in Table 5.

5.2 Performance metrics

The performance metrics used to evaluate the performance of this proposed method are accuracy, precision, sensitivity, specificity, false positive rate (FPR), F1-score and error. Accuracy is a performance metric, which is used to estimate the correctness of the classification. Precision is denoted as the fraction of recognized instances, which are significant, and sensitivity is denoted as the fraction of significant instances, which are retrieved. Specificity is a measure, which is used to evaluate the proportion of actual negatives that are correctly identified. The ratio of number of negatives that are incorrectly represented as positive to the total number of actual negatives provides the false positive rate (FPR). The F1-score is the harmonic average of the precision and recall. Error is used to evaluate the misclassification percentage. The formulas used for the performance evaluation are given in Equations (30)(31) (32)(33)(34)(35).

\[
\text{Accuracy} = \frac{TP + TN}{\text{Total}} \quad (30)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (31)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (32)
\]

\[
F1 - \text{score} = \frac{2TP}{2TP + FP + FN} \quad (33)
\]

\[
\text{Specificity} = \frac{TN}{TN + FP} \quad (34)
\]

\[
\text{FPR} = \frac{FP}{FP + TN} \quad (35)
\]

![FIGURE 8 Comparison of sensitivity for various datasets](image-url)
\[ \text{Error} = 100 - \text{accuracy} \quad (36) \]

where TP indicates true positive, TN indicates true negative, FP indicates false positive, and FN indicates false negative.

5.3 | Performance evaluation

The methods that are taken for comparison are SVR, DBN, and RNN. The brief description for this three techniques is explained below:

**DBN:** DBN stands for DBN, which is a famous deep learning architecture. The layered structure in this algorithm is mainly utilised to learn the hidden structure of the patterns, and it is used to extract the important parameters from the data. Moreover, number of RAMs are stacked in this structure; so, it learns the high level representations, which are valuable for music emotion recognition. The total nodes found in each layer is 18-30, 30-30, 30-30, and 30-7. Total RBM is 3, learning rate is 0.1, batch size is 84, and weight decay is 0.0002.

**SVR:** SVR are used in SVMs for regression problems. Main goal of this SVR is discover a hyperplane in a feature space like SVM. The hyperplane of SVR correctly predicts the distribution of original information while the hyperplane of SVM separates the data into two parts.

**RNN:** Neural network concept and rough set theory are basics for RNN. Feedback loops among the RNN neuron is considered as the main difference between conventional NN and RNN; so, the RNN has memory function. The learning rate of RNN is 0.1 and 0.001 as well as the initial values of weight and biases are chosen for truncated normal and empirical value. [None, 1092] is the dimension of input layer in which 1092 is length of the MIDI music feature matrix of the interception. [input_size, 1560] is hidden layer dimension value in which 1560 is the number of hidden layer neurons set.

![Figure 9](https://via.placeholder.com/150)

**FIGURE 9** Comparison of precision for various datasets
Finally, [1560, five] is a dimension of output layer in which the number of output music emotion label is mentioned as five.

The performance metrics are evaluated for the proposed and existing methods, and the proposed method's performance values are tabulated in Table 6.

The proposed technique has the highest accuracy of 98.79% when using songs from ISMIR2012 dataset. From the same dataset when using karaoke of the songs, our proposed method attains the accuracy of 97.4%. When using the lyrics from ISMIR2012 dataset, the highest accuracy is achieved for our proposed method, which is 97.23%. On comparing the three datasets, ISMIR2012 dataset achieved the highest accuracy for songs, karaoke and lyrics. Also, when considering songs, karaoke and lyrics, the accuracy is high for songs. When using the NJU_V1 dataset, the highest accuracy for songs, karaoke and lyrics are 97.88%, 96.82% and 95.76% respectively. When using the developed Hindi dataset, the highest accuracy for songs, karaoke and lyrics are 97.57%, 95.63% and 94.17% respectively. Similarly for remaining two datasets, the proposed approach SVR_KMBSO has attained higher accuracy result most particularly for songs.

Figures 4–6 represent the confusion matrices of song, karaoke and lyrics for ISMIR2012, NJU_V1, and Hindi datasets. From the above confusion matrix, the values of TP, FP, FN and TN for the four classes are evaluated based on the correct and incorrect classification. When considering the songs, the TP values of angry, happy, relax and sad are 125, 149, 147 and 150 respectively. The FP values of angry, happy, relax and sad are 2, 2, 2 and 1, respectively. The FN values of angry, happy, relax and sad are 2, 1, 2 and 2, respectively. The TN values of angry, happy, relax and sad are 449, 426, 427 and 425, respectively. When considering the karaoke, the TP values of angry, happy, relax and sad are 123, 146, 145 and 149, respectively. The FP values of angry, happy, relax and sad are 3, 5, 3 and 4, respectively. The FN values of angry, happy, relax and sad are 4, 4, 4 and 3, respectively. The TN values of angry,

![Comparison graph of F1-score for various datasets](image)

**FIGURE 10** Comparison graph of F1-score for various datasets
happy, relax and sad are 448, 423, 426 and 422, respectively. When considering the lyrics, the TP values of angry, happy, relax and sad are 123, 148, 144 and 147, respectively. The FP values of angry, happy, relax and sad are 4, 5, 3 and 4, respectively. The FN values of angry, happy, relax and sad are 4, 2, 5 and 5, respectively. The TN values of angry, happy, relax and sad are 447, 423, 426 and 422, respectively.

When considering the songs, the TP values of angry, happy, relax and sad are 70, 104, 98 and 97, respectively. When considering the karaoke, the TP values of angry, happy, relax and sad are 68, 103, 98 and 96, respectively. When considering the lyrics, the TP values of angry, happy, relax and sad are 68, 102, 97 and 94, respectively. The emotions that are correctly classified are comes under the TP class. The presence of modified BSO has improved the recognition rate of SVR; due to this, the accuracy rate is also gets improved.

The overall accuracy attained by proposed approach for song, karaoke, and lyrics are found to be 97.57%, 95.63%, and 94.17% respectively. The accuracy value attained by proposed SVR, KMBSO for song in Hindi dataset is 98.21 (angry), 97.5 (happy), 97% (relax), and 97.67 (sad).

The accuracy, precision, sensitivity, F1-score, specificity, error, and FPR comparison of the proposed method for various datasets using existing SVR, DBN and RNN classifiers are shown in Figures.

Figure 7 illustrates the accuracy comparison of the proposed method and the existing methods for various datasets using SVR, DBN and RNN classifiers. The above graph represents that the proposed technique has the highest accuracy of 98.79% when using songs from ISMIR2012 dataset. From the same dataset when using karaoke of the songs, our proposed method attains the accuracy of 97.4%. When using the lyrics from ISMIR2012 dataset, the highest accuracy is achieved for our proposed method, which is 97.23%. On comparing the three datasets, ISMIR2012 dataset achieved the highest accuracy for songs, karaoke and lyrics. Also, when considering songs, karaoke and

![Comparison graph of specificity for various datasets](image_url)
lyrics, the accuracy is high for songs. When comparing the classifiers, the proposed SVR_KMBSO method achieves the best accuracy for songs, karaoke as well as for lyrics. When using the NJU_V1 dataset, the highest accuracy for songs, karaoke and lyrics are 97.88%, 96.82% and 95.76%, respectively. When using the developed Hindi dataset, the highest accuracy for songs, karaoke and lyrics are 97.57%, 95.63% and 94.17%, respectively.

Figure 8 represents that the proposed method has the highest sensitivity of 98.78% when using songs from ISMIR2012 dataset. From the same dataset when using karaoke of the songs, our proposed method attains the sensitivity of 97.38%. When comparing the classifiers, the proposed SVR_KMBSO method achieves the best sensitivity for songs, karaoke as well as for lyrics.

From the same dataset when using karaoke of the songs, our proposed method attains the precision of 97.42% (Figure 9). When using the lyrics from ISMIR2012 dataset, the highest precision is achieved for our proposed method, which is 97.22%. On comparing the three datasets, ISMIR2012 dataset achieved the highest value of precision for songs, karaoke and lyrics. Also, when considering songs, karaoke and lyrics, the precision value is high for songs.

The proposed method has the highest F1-score of 98.77% (Figure 10) when using songs from ISMIR2012 dataset. From the same dataset when using karaoke of the songs, our proposed method attains the F1-score of 97.4%. When using the lyrics from ISMIR2012 dataset, the highest F1-score is achieved for our proposed method which is 97.22%.

The specificity graph in Figure 11 represents that the proposed method has the highest specificity of 99.6% when using songs. On comparing the three datasets, ISMIR2012 dataset achieved the highest value of specificity for songs, karaoke and lyrics. When using the NJU_V1 dataset, the highest specificity values for songs, karaoke and lyrics are 99.29%, 98.95% and 98.58%, respectively. When using the developed Hindi dataset, the highest specificity values for

**FIGURE 12** Comparison graph of false positive rate (FPR) for various datasets
songs, karaoke and lyrics are 99.2%, 98.55% and 98.05%, respectively.

Figure 12 represents that the proposed method has the lowest FPR of 0.40% when using songs from ISMIR2012 dataset. From the same dataset when using karaoke of the songs, our proposed method attains the FPR of 0.87%. When using the lyrics from ISMIR2012 dataset, the lowest FPR is achieved for our proposed method which is 0.92%. On comparing the three datasets, ISMIR2012 dataset achieved the lowest FPR for songs, karaoke and lyrics.

From the comparison shown in Figure 13, we can conclude that the proposed method has achieved the lowest error of 1.21% for songs from ISMIR2012 dataset. For three datasets, the error rate attained by proposed SVR_KMBSO is found minimum than other existing techniques. This is because the SAE in proposed approach extracts only the high-level features, and it avoids the low-level features. Elimination of low-level features reveals more dimension, which automatically reduces the burden of classification process. Therefore, proposed SVR_KMBSO has attained less error rate.

The training and testing accuracy of ISMIR2012 dataset, NJU_V1 dataset and Hindi dataset with respect to songs, karaoke and lyrics are shown in Figures 14–16 and the Tables 7–9 shows the Figures 14–16 in detail. Total files in these three datasets are equally divided for training and testing. ISMIR2012 contains 2850 files, among that 1425 are used for testing, and 1425 are for training. NJU-V1 datasets contains 775 files, whereas 388 are used for training and 387 are for testing. Third dataset is Hindi dataset which contains 925 files, among that 465 is used for training and 460 is for testing.

Training accuracy is identified by estimating the classification accuracy of a model with the same data in which the model is trained. Testing accuracy is estimated by introducing two completely different datasets. One is to train the model which is

![Comparison graph of error for various datasets](image-url)
called as training data and the other is to calculate the classification accuracy which is called as testing data. If the number of files increased, then the training and testing accuracy also gets increased. While comparing with the testing accuracy, the accuracy attained for training is found greater for all sets including songs, karaoke and lyrics. When considering the training accuracy of songs and karaoke for 1425 samples, it is 2% higher for the proposed method than SVR method, 6% (song), and 7% (lyrics) higher than DBN approach and 8% higher than RNN method. When considering the training accuracy of lyrics for 1425 samples, it is 5% higher for the proposed method than the SVR method, 6% higher than DBN approach and 7% higher than RNN method. When considering the testing accuracy of karaoke and lyrics for 1140 samples, the proposed method’s accuracy is nearly 4% (karaoke), and 5% (lyrics) higher than SVR method, 12% (karaoke), and 14% (lyrics) higher than DBN method and 14% (karaoke), and 13% (lyrics) higher than RNN method. Normally, the training accuracy is higher than the testing accuracy, but in some cases the training accuracy is found less than the testing accuracy, this is because the features that are selected for emotion recognition cause under-fitting during learning process. To overcome such issue, the KMBSO algorithm is hybrid with SVR instead of BP. This hybrid technique attains high training accuracy than the testing accuracy.

Table 8 shows the training and testing accuracy of NJU_V1 dataset with respect to songs, karaoke and lyrics. The training and testing accuracy of proposed SVR_KMBSO is compared with three existing techniques they are SVR, DBN, and RNN.
Among these all techniques the accuracy attained by proposed SVR_KMBSO is found higher that other techniques. From this result we can conclude that our proposed SVR_KMBSO attains better accuracy result for songs rather than for lyrics and karaoke. This is because song based emotion recognition have inherit the benefits of both text dependent and text independent features. Due to this the accuracy of song based music recognition is found high.

Abbreviations: DBN, deep belief network; SVR_KMBSO, K-Medoid-based brain storm optimisation-based support vector regression; SVR, support vector regression, RNN, recurrent neural network.

Table 9 shows the training and testing accuracy of Hindi dataset with respect to songs, karaoke and lyrics. When compared to testing accuracy, the training accuracy is highly achieved for all sets including songs, karaoke and lyrics. When considering the training accuracy of songs for 465 samples, it is 3% higher for the proposed method than SVR method, 7% higher than DBN approach and 6% higher than RNN method.

The accuracy result of proposed approach is compared with different existing techniques, and the comparison results are shown in Table 10. From this comparison it is clear that among all existing techniques, the performance of proposed (SVR_KMBSO) is found better because it includes a modified BSO algorithm along with SVR to improve the recognition result. Finally, this hybridisation process has achieved better accuracy result with less error rate. The computational complexity achieved by proposed SVR_KMBSO is found less

TABLE 7 Training and testing accuracy of ISMIR2012 dataset

| No of Files | Training accuracy | Testing accuracy |
|-------------|-------------------|------------------|
|             | 285   | 570   | 855   | 1140  | 1425  | 285   | 570   | 855   | 1140  | 1425  |
| **Song**    |       |       |       |       |       |       |       |       |       |       |
| SVR_KMBSO   | 89.52 | 92.58 | 94.96 | 97.85 | 99.88 | 87.63 | 91.85 | 92.28 | 95.37 | 98    |
| SVR         | 80.15 | 82.68 | 89.96 | 93.25 | 97.8  | 79.22 | 81    | 87.956| 91.26 | 96.37 |
| DBN         | 73.54 | 76.9  | 79.51 | 85.57 | 93.6  | 72.8  | 75.14 | 78.27 | 84.42 | 92.9  |
| RNN         | 71.65 | 75.28 | 76.15 | 85.77 | 91.82 | 70.8  | 74.96 | 75.57 | 84.13 | 90.78 |
| **Karaoke** |       |       |       |       |       |       |       |       |       |       |
| SVR_KMBSO   | 86.83 | 91.145| 92.21 | 96.82 | 99.35 | 85.58 | 90.8  | 91.56 | 95.89 | 98.42 |
| SVR         | 77.537| 81.43 | 87.6  | 92.48 | 97.656| 76.81 | 79.56 | 86.65 | 91.31 | 95.6  |
| DBN         | 71.58 | 73.25 | 78.35 | 84.84 | 92.35 | 70.25 | 72.659| 77.128| 83.45 | 90.9  |
| RNN         | 70.28 | 71.95 | 75.15 | 83.45 | 91.69 | 69.53 | 71.56 | 74.58 | 81.61 | 89.92 |
| **Lyrics**  |       |       |       |       |       |       |       |       |       |       |
| SVR_KMBSO   | 85.28 | 90.5  | 92.39 | 95.38 | 98.2  | 84.53 | 88.8  | 90.9  | 94.59 | 97.583|
| SVR         | 75.58 | 78.51 | 84.456| 90.85 | 93.56 | 74.12 | 77.56 | 83.6  | 89.23 | 92.26 |
| DBN         | 70.81 | 73.25 | 77.215| 82.24 | 92.23 | 68.35 | 71.69 | 76.22 | 80.95 | 90.869|
| RNN         | 69.58 | 71.35 | 75.5  | 82.94 | 91.6  | 68.3  | 70.69 | 73.44 | 81.66 | 89.56 |

Abbreviations: DBN, deep belief network; SVR_KMBSO, K-Medoid-based brain storm optimisation-based support vector regression; SVR, support vector regression, RNN, recurrent neural network.
than other existing techniques as it use KMBSO algorithm for parameter optimisation. This parameter optimisation may avoid the overfitting meanwhile it maximise the recognition accuracy. But the existing algorithms cause solutions to fall within local optimum; therefore, the computational complexity of existing algorithms is high.

### 6 CONCLUSION

Recently, the machine learning is gaining a prominent role in each field due to its high performance. Music is automatically integrated within the everyday life of all human beings. It harmonises the sensation of the listener; moreover, it also reveals the emotion. Till now, a number of techniques exist for music mood recognition, but none of them provide better accuracy with less computational complexity. An efficient supervised framework with autoencoder-based optimised SVR model is developed for the MER. In feature extraction process, the features are extracted from songs, karaoke and lyrics as text-dependent and non-text-dependent features. For the high level, feature representation stacked autoencoder is used with two hidden layers. KMBSO-based support vector regression model is utilised for the emotion classification. The optimal parameters of the SVR are selected by the KMBSO algorithm. So, the performance of SVR gets enhanced than the conventional support vector regression. Three datasets are used in this method for performance evaluation; among that, two datasets (ISMIR2012, NJU_V1) contain English songs and one (Hindi dataset) contains Hindi songs. The training and testing
| Reference          | Song                  | Features                                      | Classification Technique | Emotions                  | Dataset                              | Performance metrics                                                                 |
|--------------------|-----------------------|-----------------------------------------------|--------------------------|----------------------------|--------------------------------------|--------------------------------------------------------------------------------------|
| Medina et al., 2020 [35] | 1802 songs, Length 45 s | Each song carries 260 low-level features      | Multi-layer perceptron, SVM, RF | All emotions               | MediaEval                           | F-measure [valence 0.77, arousal 0.72], RMSE 0.11                                   |
| Gao et al., 2020 [28]    | 1744 songs, length 45 s | Audio features and sequential features        | CRNN                     | All emotions               | MediaEval, scratch-music dataset     | Precision [0.824], recall [0.819], and F1-score [0.821]                             |
| Sarkar et al., 2019 [26] | Sound track [360 songs, length 30 s] | Different spectral and time domain features, MFCC and LPC features | CNN built around VGGNet  | Happy, anger, tender, and sad | Soundtracks, Bi-modal, MER_traffic   | F-measure [MER_traffic 82.95 ± 1.42], Accuracy [67.71 ± 3.63 (sound track), 77.82 ± 4.06 (bimodal)] |
| Liu et al., 2019 [27]    | 1802 songs, length 1 s | Harmonic, features and percussive features    | RNN                      | Excited, calm, joy, sad   | Database for emotional analysis of music dataset | Accuracy [93.9 (excite), 69.32 (calm), 80.83 (joy), 92.97 (sad)]                       |
| Chaudhary et al., 2019 [23] | 1000 songs, length 30 s | Timbral feature, time-frequency energy distribution, rhythmic features | Hash-tag graph            | Anger, disgust, sad, happy, fear, and surprise | Online music database (Hindi songs) | Specificity [96.9], precision [81.3], recall [87.5], F-measure [84.33], accuracy [95.7], RMSE [0.5686], G-mean [92] |
| Sarkar et al., 2018 [22] | 360 audio clips, length 10–30 s | Low-level time domain and spectral features   | K-means and agglomerative clustering | Angry, fear, happy, sad and tender | Sound tracks dataset | Accuracy [63.63 (agglomerative), 59.09 (k-means)]                                    |
| Hu et al., 2017 [25]     | 5296 songs, length 120 s | Audio and lyric features                      | SVM and KNN               | Angry, excited, glad, calm, sad, and gloomy, romance | Million song dataset | Accuracy | Audio [64.6], lyric [63.5] |

Abbreviations: CNN, convolutional neural network; CRNN, convolutional-recurrent neural network; KNN, k-nearest neighbour; MER, music emotion recognition; RMS, root mean square error; RNN, Recurrent neural network.
accuracy for the three different datasets (ISMI2012, NJU_V1, and Hindi dataset) are evaluated. The experimental results are evaluated and compared with the existing classifiers including SVR, DBN and RNN. When compared with the existing classifiers, our proposed method SVR_KMBSO achieved a high accuracy of 98.79% for songs by using ISMI2012, NJU_V1, and Hindi dataset. Mood recognition performed by this method is found to be much better than other existing techniques. Totally, the proposed SVR_KMBSO has attained higher performance for songs; however, the performance for lyrics and karaoke are found to be little less than the song.

REFERENCES

1. Malheiro, R., et al.: Emotionally relevant features for classification and regression of music lyrics. IEEE Trans. Affect. Comput. 9(2), 240–254. (2016)
2. Panda, R., Malheiro, R.M, Paiva, R.P.: Novel audio features for music emotion recognition. IEEE Trans. Affect. Comput. 1, 1–1 (2018)
3. Lee, J., et al.: Investigating relation of music data: emotion and audio signals. In:Computer Science and its Applications, pp. 251–256.Springer, Berlin (2015)
4. Wang, J.-C., et al.: The acoustic emotion Gaussians model for emotion-based music annotation and retrieval. In: Proceedings of the 20th ACM International Conference on Multimedia, pp. 89–98 (2012)
5. Chen, Y.-A., et al.: Linear regression-based adaptation of music emotion recognition models for personalization. In: Acoustics, speech and Signal Processing (ICASSP), pp. 2149–2153.IEEE (2014)
6. Chen, S.-H., et al.: Music emotion recognition using deep Gaussian process. In: Signal and Information Processing Association Annual Summit (APSIPA), Asia-Pacific, pp. 495–498.IEEE (2015)
7. Chiang, W.-C., Wang, J.-S., Hsu, Y.-L.: A music emotion recognition algorithm with hierarchical SVM based classifiers. In: 2014 International Symposium on Computer, Consumer and Control (ISCCC), pp. 1249–1252.IEEE (2014)
8. Chen, Y.-A., et al.: Component tying for mixture model adaptation in personalisation of music emotion recognition. IEEE/ACM Trans. Audio, Speech, Language Process. 25(7), 1409–1420 (2017)
9. Subramanian, R., et al.: ASCERTAIN: emotion and personality recognition using commercial sensors. IEEE Trans. Affect. Comput. 9(2), 147–160 (2016)
10. An, Y., Sun, S, Wang, S.: Naive Bayes classifiers for music emotion classification based on lyrics. In: 201716th International Conference onComputer and Information Science (ICIS), pp. 635–638.IEEE/ACIS IEEE (2017)
11. Markov, K, Matsui, T.: Music genre and emotion recognition using Gaussian processes. IEEE Access. 2, 688–697 (2014)
12. Barthet, M., Fazekas, G, Sandler, M.: Music Emotion recognition: from content-to context-based models. In: In International Symposium on Computer Music Modeling and Retrieval, pp. 228–252.Springer Berlin Heidelberg (2012)
13. Hu, X, Yang, Y.-H.: Cross-dataset and cross-cultural music mood prediction: a case on Western and Chinese Pop songs. IEEE Trans. Affect. Comput. 8(2), 228–240 (2017)
14. Zhang, J.L., et al.: Feature selection and feature learning in arousal dimension of music emotion by using shrinkage methods. Multimed Syst. 23(2), 251–26 (2017)
15. Yang, X., Dong, Y., Li, J.: Review of data features-based music emotion recognition methods. Multimed Syst. 24(4), 365–389 (2018)
16. Baniya, B, Lee, J.: Rough set-based approach for automatic emotion classification of music. J. Inform. Process. Syst. 13(2), 400–416 (2017)
17. Ren, J-M., Wu, M-J, Jiang, J-S.R.: Automatic music mood classification based on timbre and modulation features. IEEE Trans. Affect. Comput. 3, 236–246 (2015)
18. Yang, Y-H, Chen, H.H.: Ranking-based emotion recognition for music organization and retrieval. IEEE Trans. Audio Speech Lang. Process. 19(4), 762–774 (2011)
19. Patra, B., Das, D, SivajiBandopadhyay: Labelling data and developing supervised framework for Hindi music mood analysis. J. Intell. Inf. Syst. 48(3), 633–651 (2017)
20. Patra, B., Das, D, SivajiBandopadhyay: Multimodal mood classification of Hindi and Western songs. J. Intell. Inf. Syst. 51(3), 1–18 (2018)
21. Dong, Y., et al.: Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for emotion recognition. IEEE Trans. Multimed. 21(12), 3150–3163 (2019)
22. Sarkar, R., et al.: Emotion based categorisation of music using low level features and agglomerative clustering. In 2017 6th National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics (NCVPRIPG), Mandi, India, 16–19 December, 2017, Revised Selected Papers No. 6, pp. 506–516.Springer Singapore (2018)
23. Chaudhary, D., Singh, N.P, Singh, S.: Automatic music emotion classification using hashtag graph. Int. J. Speech Technol. 22(3), 551–561 (2019)
24. Ayata, D, Yasar, Y, Kamasak, M.E.: Emotion based music recommendation system using wearable physiological sensors. IEEE Trans. Consumer. Electron. 14(8), 1–8 (2018)
25. Hu, X., Choi, K., Downie, J.S.: A framework for evaluating multimodal music mood classification. J. Assoc. Inform. Sci. Technol. 68(2), 273–285 (2017)
26. Sarkar, R., et al.: Recognition of emotion in music based on deep convolutional neural network. Multimed. Tool Appl. 79(1), 765–783 (2020)
27. Liu, H., Fang, Y, Huang, Q.: Music emotion recognition using a variant of recurrent neural network. In: 2018 International Conference on Mathematics, Modelling, Simulation and Statistics Application (MMSSA 2018), Atlantis Press January (2019)
28. Gao, Z., et al.: A novel music emotion recognition model for scratch-generated music. In: 2020 International Wireless Communications and Mobile computing (IWCMC), pp. 1794–1799. IEEE (June 2020)
29. Kim, E., Shin, J.W.: DNN-based emotion recognition based on bottleneck acoustic features and lexical features. IEEE Trans. Multimedia. 11(4), 670–682 (2009)
30. Bandhakavi, A., et al.: Lexicon based feature extraction for emotion text classification. Pattern Recogn. Lett. 93, 133–142 (2017)
31. Lee, C.H., et al.: Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. IEEE Trans. Multimedia. 11(4), 670–682 (2009)
32. Lee, C.H., et al.: Automatic music genre classification using modulation spectral contrast feature. In: IEEE International Conference on Multimedia and Expo IEEE, pp. 204–207 (July 2007)
33. Zhang, J., et al.: Emotion recognition using time-frequency analysis of EEG signals and machine learning. In: 2019 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 404–409.IEEE (2019)
34. Zhou, P., et al.: Learning compact and discriminative stacked autoencoder for hyperspectral image classification. IEEE Trans. Geosci. Rem. Sens. 57(7), 4823–4833 (2019)
35. Medina, Y.O., Beltrán, J.R., Baldassarri, S.: Emotional classification of music using neural networks with the MediaEval dataset. In: Personal and Ubiquitous Computing, pp. 1–13 (2020)

How to cite this article: Agarwal G, Om H. An efficient supervised framework for music mood recognition using autoencoder-based optimised support vector regression model. IET Signal Process. 2021;15:98–121. https://doi.org/10.1049/sil2.12015