DOI: 10.6977/IJoSI.202508_9(4).0006

Decoding Marathi emotions: Enhanced speech emotion recognition through deep belief network-support vector machine integration

Varsha Nilesh Gaikwad[1]*, Rahul Kumar Budania[2]

[1]Department of Electronics and Telecommunication Engineering, School of Engineering, RMD Sinhgad Technical Institute, Pune, Maharashtra, India

[2]Department of Electronics and Communication Engineering, Institute of Engineering, Shri JJT University, Jhunjhunu, Rajasthan, India

*Corresponding author E-mail: nvarsha29619@gmail.com

(Received 29 October 2024; Final version received 9 December 2024; Accepted 12 December 2024)

Abstract

Speech emotion recognition in Marathi presents considerable hurdles due to the language's distinct grammatical and emotional characteristics. This paper presents a robust methodology for classifying emotions in Marathi speech utilizing advanced signal processing, feature extraction, and machine learning techniques. The method entails collecting diverse Marathi speech samples and using pre-processing steps such as pre-emphasis and voice activity detection to improve signal quality. Speech signals are segmented using the Hamming window to reduce discontinuities, and features such as Mel-frequency cepstral coefficients, pitch, intensity, and spectral properties are retrieved. For classification, an attentive deep belief network is paired with a support vector machine, which uses attention techniques and batch normalization to improve performance and reduce overfitting. The suggested approach surpasses existing models, with 98% accuracy, 98% F1-score, 99% specificity, 99% sensitivity, 98% precision, and 98% recall.

Keywords: Speech Emotion Recognition, Voice Activity Detection, Mel-Frequency Cepstral Coefficient, Deep Belief Network, Support Vector Machine

1. Introduction

User interfaces are growing more complicated, with voice processing technology allowing users to communicate without physically using a keyboard (Chaudhari et al., 2023). Speech is an important type of human-to-human communication that provides emotional and psychological information. Speech processing provides sound qualities and characteristics that can be used to extract meaningful information (Papala et al., 2023). Speech emotion classification is not only at the core of human life and action, as most scientific and psychiatric endeavors have shown, but it can also be studied using the computing tools of today's modernist conception of science. One unanswered topic, nevertheless, is how the application of machine learning techniques to the study of the typical, human, and empirically observed dynamics of emotion classification has changed, evolved, or expanded in scope (Akinpelu & Viriri, 2024). However, identifying emotions from speech remains challenging due to the range of expressions, even for the same feeling (Lieskovská et al., 2021). Joy, fury, fear, and sorrow have similar acoustic features, such as voice volume, pitch, and the number of times their speech meets the zero axis (Madanian et al., 2023). This issue stems from the recognition of these two sets of emotions, which we extract directly from speech signals or text, and the feature set used for emotion detection (Hammed & George, 2023). Acoustic elements of speech, such as pitch, intensity, and volume, can also be deceptive when considered alone (Kaur & Singh, 2023). People employ speech signal features and speech semantics to communicate their emotions in everyday situations, emphasizing the significance of extracting emotions from both acoustic and semantic variables before concluding the underlying emotions in a speech signal (Zaidi et al., 2023).

Artificial intelligence (AI) advancements have improved the comfort and convenience of human-computer contact (Yang et al., 2024). The next wave of AI development will focus on enhancing speech emotion recognition (SER), which has both theoretical and practical ramifications (Harhare & Shah, 2021). Feature extraction is critical in speech signal processing, and hand-designed features have been used for SER (Bachate et al., 2022). The spectral feature, which considers both the frequency and time axes, has gained prominence in recent years. There are several challenges with the traditional method of identifying emotion from speech utterances. Many of the current methods, including the support vector machine (SVM), hidden Markov model, and Gaussian mixture model, rely on automated speech recognition, which is highly dependent on dataset manipulation. Any changes may necessitate reconstructing the entire model. It is impossible to categorize emotion lightly because it contains important information that has the power to either make or break a person's personality. To avoid some of these issues with the conventional method (Akçay & Oğuz, 2020). However, these manual abilities are limited and cannot adequately portray emotions in speech. To solve this, neural networks have proposed a solution that incorporates deep learning into the model-building process (Alam Monisha & Sultana, 2022).

Deep learning features extract specialized feature representations from big learning problems, which reduce the incompleteness caused by artificially created features (Padman & Magare, 2022). The standardized pre-trained (Chai et al., 2021) model is typically used to address the issue of the inadequate training dataset in transfer learning, a fundamental area of deep learning that has demonstrated effectiveness in a variety of computer vision-related applications, including emotion identification (Li et al., 2021). It is a deep convolutional neural network (DCNN) subdivision. Due to its inherent capacity to extract speech features from speech signals distinctively and efficiently, the DCNN application to emotion categorization gained prominence (Oh & Kim, 2022). Researchers are constantly on the lookout for new DCNN techniques (Byun & Lee, 2021) that can produce more noticeable results, but the current findings have exposed the long-standing issues of inadequate label datasets for the classification of speech emotion and a high level of parameterization of the field. As a result, there is a need to develop a dependable method for automatic speech detection.

2. Literature Survey

Section 2 presents an overview of the current research in SER, summarizing and discussing key literature. SER is critical for understanding human emotional behavior and relies on identifying distinguishing traits. Alluhaidan et al. (2023) enhanced SER system performance by combining Mel-frequency cepstral coefficients (MFCCs) and time-domain features. To create the SER model, a convolutional neural network (CNN) was fed the suggested hybrid features. The current work limits the acquisition of high-level acoustic information crucial for accuracy because it does not compare SER approaches across datasets and does not include recurrent neural networks. Kawade & Jagtap (2024) employed a DCNN and multiple acoustic features and proposed a cross-corpus SER (CCSER) for an Indian corpus. For feature selection, Fire Hawk-based optimization reduces computational complexity and enhances feature distinctiveness. Better correlation, greater feature representation, and a more accurate description of the speech signal's timbre, intonation, and pitch variation are all provided by the DCNN method. However, its capacity to generalize across a variety of emotional circumstances is diminished by its lack of domain adaptability and inadequate global and local acoustic properties. Bhangale & Kothandaraman (2023) displayed the acoustic feature set using the following methods to increase the feature distinctiveness: MFCC, linear prediction cepstral coefficients (LPCCs), wavelet packet transform (WPT), zero crossing rate (ZCR), spectrum centroid, spectral roll-off, spectral kurtosis, root mean square (RMS), pitch, jitter, and shimmer. In addition, a lightweight, compact one-dimensional DCNN is employed to capture the spoken emotion signal's long-term relationships and reduce computational complexity. The system is not resilient under CCSER under different noise settings and suffers from class imbalance as a result of unequal dataset training. Farooq et al. (2020) conducted a project to improve SER by employing a DCNN to classify emotions during human-machine interaction accurately. The DCNN extracts features from complex speech-emotional datasets using a correlation-based feature selection method. The approach obtains 95.10% accuracy in speaker-dependent SER tests with four publicly available datasets: Emo-DB, SAVEE, IEMOCAP, and RAVDESS. This is especially essential in audio conferencing, where traditional machine learning methods are less trustworthy due to noise sensitivity and accent fluctuations. Sonawane & Kulkarni (2020) used a deep learning strategy for emotion speech detection that employs a multilayer CNN and a simple K-nearest neighbor classifier, which has been shown to outperform the current MFCC method in real-time testing on the YouTube database. This technology is critical for SER in real-time applications such as human behavior assessment, human–robot interaction, virtual reality, and emergency rooms. Sajjad & Kwon (2020) created a new framework for SER that selects sequence segments based on cluster-level similarity measures. The short-time Fourier transform technique transforms the sequence into a spectrogram, which is then input into a CNN model for feature extraction. The CNN features are normalized for accurate recognition and input into bidirectional long short-term memory for emotion recognition. The system is evaluated using typical datasets to enhance recognition accuracy and processing time.

The contributions of this work are as follows:

- The study focuses on creating a diverse dataset of Marathi voice samples to capture the unique grammatical and emotional characteristics of the Marathi language.
- It employs advanced feature extraction techniques, including pre-emphasis; voice activity detection (VAD); and acoustic, prosodic, and spectral features, to capture nuanced emotional traits in Marathi speech.
- The study also develops a hybrid classification model combining a deep belief network (DBN) and an SVM, enhancing feature representation and emotion recognition accuracy.

3. Proposed Methodology

SER in Marathi is a significant difficulty due to the language's unique grammatical and emotional characteristics. To address these challenges, a complex methodology was developed for improving emotion classification accuracy in Marathi speech samples. The dataset compilation included specifics about the Marathi voice samples, such as their source, size, emotional categories, and demographic diversity, alongside the procedures for collection and annotation. For pre-processing, the exact parameters for pre-emphasis filters improved the speech signal by removing low-frequency noise. VAD was then utilized to determine the beginning and conclusion of speech by combining temporal and frequency domain approaches. The speech signal was divided into smaller frames using a Hamming window to reduce disruptions at frame boundaries, which must be clearly outlined. Feature extraction processes detailed the computation of MFCCs, including the number of coefficients and window settings, as well as the calculation methods for prosodic features such as pitch, intensity, and duration, and spectral features such as spectral centroid and zero-crossing rate, spectral bandwidth, and spectral roll-off, provide additional information about the frequency distribution and periodicity of the speech signal. The model design specified the architecture of the attentive DBN, including layer configurations, hidden units, activation functions, and the attention mechanism, as well as the kernel type and hyperparameters for the SVM. Regularization techniques, such as batch normalization, included parameter settings, and the training process were described in terms of optimizers, learning rate, epochs, and stopping criteria. Evaluation methods outlined the performance metrics, validation strategies, and comparison benchmarks. This methodology aims to overcome the limits of existing SER methods by increasing the feature representation of Marathi speech and employing modern machine learning algorithms for more accurate emotion recognition (Fig. 1).

3.1. Pre-emphasis

A pre-emphasis filter was used to boost the high-frequency components of an audio signal. This step was essential for capturing the characteristics of the input speech samples. In our method, we started by eliminating noise from the input samples, which helped with feature extraction.

3.2. VAD

The VAD is a technique for determining whether speech is present in each frame of a noisy signal. It consists of two processing stages: gathering features from noisy signals to discriminate between speech and noise and applying a detection approach to these data. This article examines the extraction of features and the performance of VAD algorithms (Fig. 2). Speech detection has poor temporal resolution since it is usually divided into shorter frames rather than deciding for each sample: Equations (1-3).

$$xl = [x(lL - N + 1),\dots, x\,(lL - 1), x\,(lL)\,]^2 \tag{1}$$

$$H_1\text{: } x(l) = b(l) + s(l) \tag{2}$$

$$H_0\text{: } x(l) = b(l) \tag{3}$$

The noisy frame can be assumed to be a combination of speech components (s [L] and noise (b [L]) or simply noise. The decision for one hypothesis, Equations (4 and 5).

$$VAD_{ftr}\left(n,l\right) = \begin{cases} 1,\ When\ H_1\ is\ accepted \\ 0\ When\ H_0\ is\ accepted \end{cases} \tag{4}$$

$$VAD_{ftr}\left(n,l\right) = \begin{cases} 1,\ where\ ftr(x(l) > n \\ 0,\ where\ ftr(x(l) \le n \end{cases} \tag{5}$$

3.3. Acoustic Feature

Acoustic aspects of a voice signal describe its physical characteristics in terms of frequency,

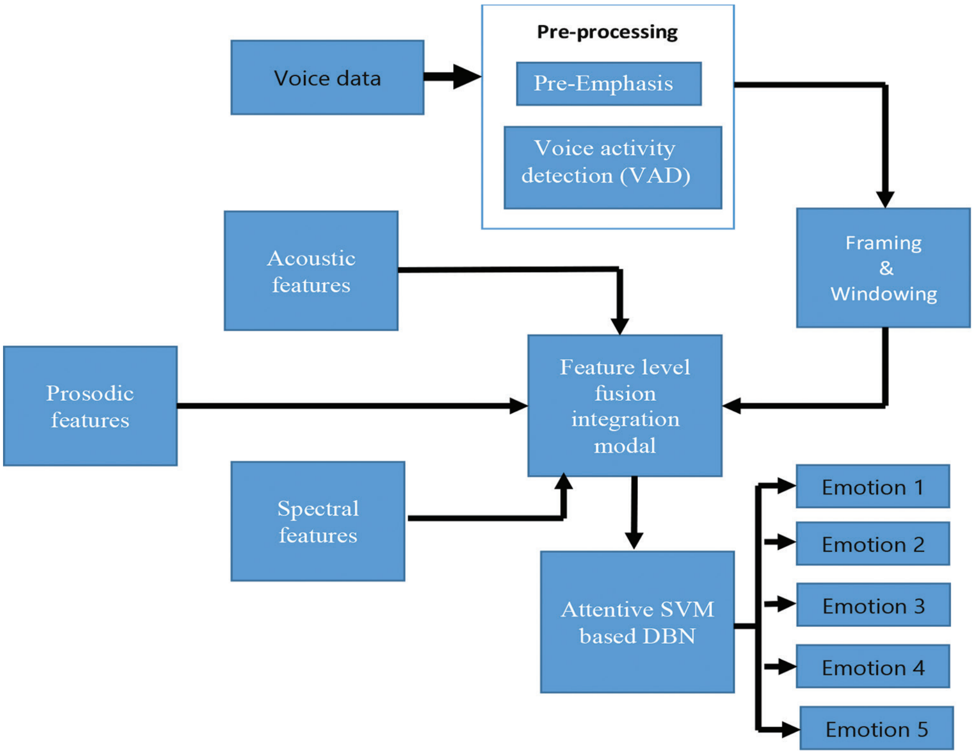

Fig. 1. Block diagram of the proposed methodology
Abbreviations: DBN: Deep belief network; SVM: Support vector machine

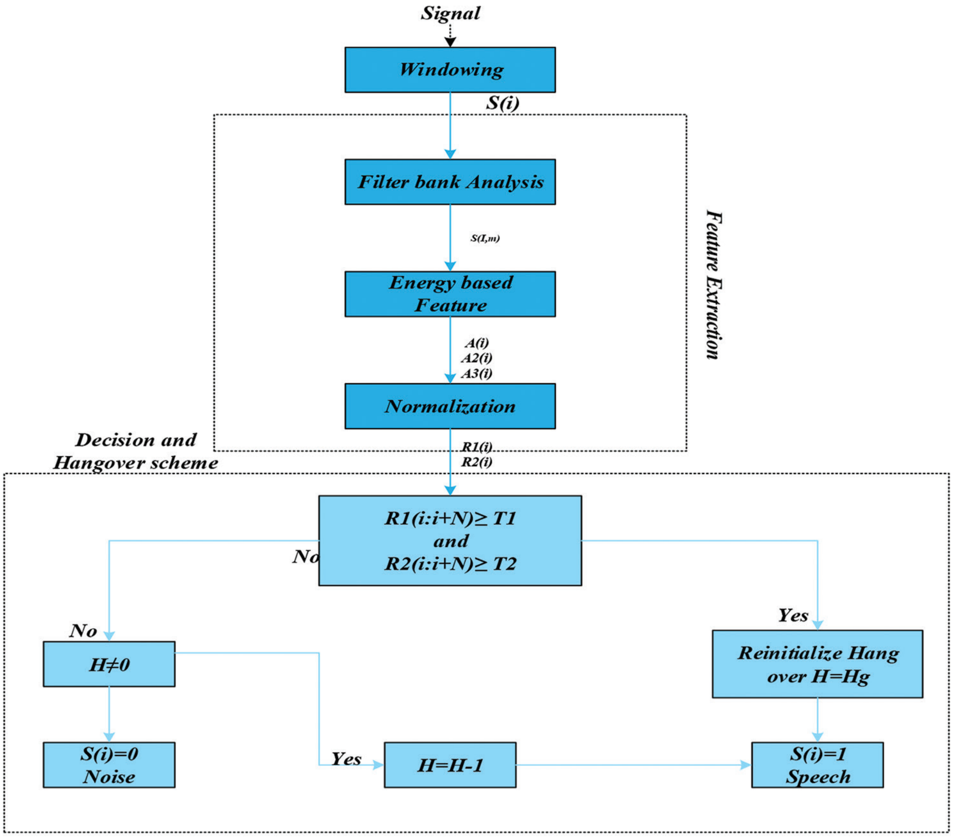

Fig. 2. Block diagram of voice activity detection

amplitude, and volume. The proposed acoustic feature set consisted of various spectral features, time-domain features, and voice quality factors that describe speech emotion. The acoustics features extracted are MFCC, LPCC, WPT, ZCR, RMS, SK, jitter, shimmer, pitch frequency, formants, and their mean and standard deviation. To eliminate noise and disturbances, the speech stream was sent through a moving average filter before being processed into various properties.

3.3.1. MFCC feature extraction

The MFCC is a technique for extracting spectral data regarding speech and human hearing perception (Abdel-Hamid et al., 2020; Shah et al., 2021). It entails normalizing the emphasis, removing noise and disturbances in raw emotional speech, and dividing the signal into 40-ms frames with a 50% frameshift (Fig. 3). For four-second voice signals, 199 frames were generated, each with a 40-ms frame width and 50% overlapping. Equation (6) demonstrates that the nearest frequency components are combined with a single Hamming window and a sample duration of 30 ms.

$$H\left(n\right)=\left(1-\alpha\right)-\alpha\times cos\frac{2\pi n}{\left(N-1\right)},0\leq n\leq N-1 \tag{6}$$

The discrete Fourier transform converts time-domain emotion speech data into frequency-domain counterparts, revealing vocal tract characteristics. The signal was processed using Mel-frequency triangular filter banks, which provided perceptual information for speech hearing. Equations (10 and 11), which convert linear to Mel frequency and vice versa, ensure appropriate interpretation of speech-hearing perceptual information.

$$X\left(K\right)=\sum_{n=0}^{N-1}x\left(n\right)\times H\left(n\right)\times e^{-\frac{j2\pi nk}{N}},0\leq n,k\leq N-1 \tag{7}$$

$$X_k=\frac{1}{N}\,|\,X(K)\,|^2 \tag{8}$$

$$ET_m=\sum_{k=0}^{k=1}\nabla_m\left(k\right)\times X_k;m=1,2,\ldots M \tag{9}$$

$$Mel=2595\log1+\frac{f}{700} \tag{10}$$

$$f=7010^{\frac{Mel}{2595}}-1 \tag{11}$$

The discrete cosine transform of the log-filter bank energy signal yields an L number of cepstral coefficients, as shown in Equation (12).

$$MFCC_i=\sum_{m=1}^{M}log_{10}\left(ET_m\right)\times cosj(m+0.5)\frac{\pi}{m} \quad \text{For}$$

j=1, 2… L (12)

According to earlier research, the MFCC contains 39 variables, including the speech signal's energy, 12 coefficients, and 26 derivatives, all of which are critical for distinguishing emotional speech shifts (Er, 2020; Kishor & Mohanaprasad, 2022).

3.4. Prosodic Feature

Prosodic features are components of speech that extend beyond phonetic segments, including intonation, stress, rhythm, and tempo. These qualities are essential for conveying meaning, emotion, and intent in spoken language. Understanding and analyzing prosodic features can help improve speech recognition systems, natural language processing, and communication interfaces. Prosodic features such as pitch, intensity, and duration are crucial for analyzing and interpreting vocal characteristics, especially in the context of emotional expression.

To calculate the pitch period (T_0), the following formula in Equation (13) was used,

$$T_o=\frac{1}{F_0} \tag{13}$$

Intensity assesses the loudness or vigor of the voice. Variations in intensity can reflect emotional states, with higher intensity frequently associated with anger or enthusiasm and lower intensity with melancholy or peacefulness (Equation [14]).

$$RMS=\sqrt{\frac{1}{N}\sum_{N=1}^{N}x[i^2]} \tag{14}$$

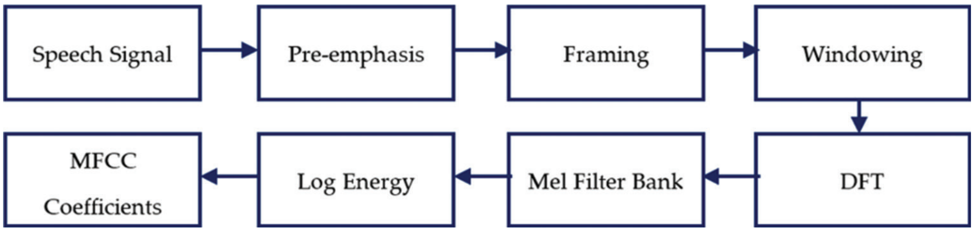

Fig. 3. Process flow of Mel-frequency cepstral coefficient feature extraction
Abbreviation: DFT: Discrete Fourier transform

Duration refers to the length of time a sound is held. The duration of spoken words and the gaps between them can reveal information about a speaker's emotional state. For example, lengthier durations and pauses may imply reluctance or reflection, whereas shorter durations may show hurry or excitement (Equation [15]).

$$Duration = N \times T_s \tag{15}$$

Where $T_N = \dfrac{1}{f_s}$ and windowing f_s is the sampling frequency.

3.5. Spectral Features

They are important in audio signal processing because they provide a complete picture of the signal's properties. Each of these characteristics contains important information about the signal's frequency distribution and periodicity, resulting in a more complete and detailed feature set.

A higher spectral centroid usually indicates a brighter sound, which is common in speech and music analysis. For example, in voice processing, the spectral centroid can aid in discriminating between distinct phonemes and emotional tones.

Spectral bandwidth measures the width of the spectrum and offers information about the range of frequencies in the signal. This function is especially important for determining the timbral properties of sound. A broader bandwidth typically suggests a richer and more complex sound, which can be critical in music genre classification and speaker recognition.

The spectral roll-off is useful for distinguishing between percussive and harmonic sounds. In music, it can aid in instrument detection, while in speech processing, it can help distinguish between voiced and unvoiced speech parts.

The ZCR is the rate at which a signal changes sign from positive to negative, or vice versa. It is a simple but powerful function for detecting the noise and roughness of an audio source. High ZCR readings frequently indicate the presence of high-frequency components, which are found in fricative sounds in speech and some musical instruments such as cymbals.

Incorporating these spectral properties into audio analytic frameworks can improve the accuracy and resilience of a wide range of applications, including speech recognition, music information retrieval, and audio categorization. These elements enabled a more comprehensive comprehension of the audio content by collecting specific information about the signal's frequency distribution and periodicity.

3.6. DBN

The DBNs are classified and feature extracted using restricted Boltzmann machines with two visible and hidden layers (Li et al., 2022). These layers are connected by weights, but nodes within the same layer are not. Backpropagation was used to train DBNs, as it is for all multilayer neural networks. In the first step, input data were used to forward-propagate restricted Boltzmann machines in each layer, and high-level abstractions were constructed by translating feature vectors to different feature spaces. V0 acted as both the first visible and input layer, whereas parameter W0 was learned from training data to rebuild the hidden and second visible layers (Arul, 2021) (Fig. 4).

A DBN is useful for estimating the posterior probability of a given feature vector. The parameters of DBN are W, b, and c. The probability of input vector v and output vector h is given below in Equation (16):

$$p(v,h) = \frac{e^{-E(v,h)}}{Z} \tag{16}$$

Where E (v, h) is the energy function (Equation [17]):

$$E(v, h) = -b^T v - c^T h - h^T Wv \tag{17}$$

The normalizing factor, Z, was calculated by adding the numerator of (16) to all conceivable h and v statuses (Equation [18]).

$$Z = \sum\nolimits_{v,h} e^{-E(v,h)} \tag{18}$$

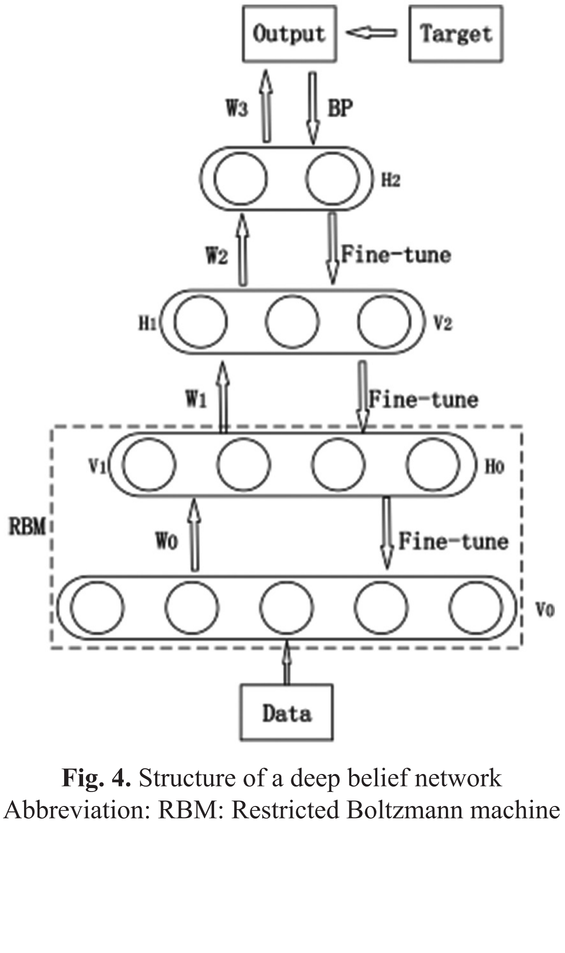

Fig. 4. Structure of a deep belief network
Abbreviation: RBM: Restricted Boltzmann machine

The DBN produced discriminative features that resemble non-linear correlations in voice samples, hence reducing the need for sophisticated feature extraction and selection. Hidden layers learned audio features from the input layer, which are subsequently used by the output layer to classify the input sample.

3.7. SVM for Emotional Classification

SVM is a nonlinear classifier that employs a kernel mapping function to convert input feature vectors into a higher-dimensional feature space (Fig. 5). To maximize discrimination, the separation plane should be advantageously positioned between the borders of two classes (Kok et al., 2024). Support vectors span the plane and reduce the number of references. The global goal is to find the equation of a hyperplane that divides *p* (Equation [19]).

$$y_i\,[(w.x_i) + b] \geq 1 \;\; \forall i = 1,2,\ldots,N \tag{19}$$

The pair (*w, b*) defines a hyperplane (Equation [20]):

$$(w.x_i) + b = 0 \tag{20}$$

This plane is known as the separating hyperplane. To identify the best-separating hyperplane, the following optimal problem in Equations (21 and 22) was considered:

$$minimize\, w(\propto) = \sum\nolimits_{i=1}^{N} \propto_i - \frac{1}{2} \sum\nolimits_{i,j,=1}^{N} \propto_i \propto_j y_i y_j K(x_i, x_j) \tag{21}$$

$$\text{Subject to } \sum\nolimits_{i=1}^{N} \propto_i y_i = 0, \propto_i \geq 0, \forall i = 1,2,\ldots,N \tag{22}$$

Proper non-linear kernels can transform non-linear classifiers into linear ones in the feature space. Some common kernel functions (Tiwari et al., 2022) are listed below in Equations (23-25):

Linear kernel:

$$K(x_i, x_j) = x_i.x_j \tag{23}$$

Polynomial kernel:

$$K(x_i, x_j) = x_i.x_j + \beta)^d \tag{24}$$

Radial basis function kernel:

$$k\left(x_i, x_j\right) = exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right) \tag{25}$$

For classification, a novel attentive DBN was paired with an SVM. This hybrid technique used the strengths of both models to improve classification performance. Attention methods were built into the DBN to dynamically weigh the contributions of various hidden units at its levels. This enabled the model to focus on salient traits that are most essential to the classification goal while ignoring irrelevant ones, resulting in increased accuracy. Regularization approaches, such as batch normalization, were also used to prevent overfitting in the model. These strategies serve to stabilize the learning process and increase the model's generalization capacity, ensuring that it performs well on both training and unseen data. This strategy was especially effective for complex classification tasks since it combines attentive processes with robust regularization methods.

4. Results and Discussion

The implementation was conducted using the Python programming language on a Windows 7 (64-bit) operating system with an Intel Pentium CPU and 8 GB of memory.

4.1. Data Description

The trials were carried out with the EMODB speech emotion database, which contains 535 utterances of seven emotions from ten German professional actors (Abdusalomov et al., 2023), and the RAVDESS emotional speech dataset, which contains 1440 samples from 24 actors. The RAVDESS dataset contains five emotions: anger, none, happiness, calmness, and fear (Abdusalomov et al., 2023). The original EMODB database samples are down-sampled to 16 kHz, yielding four-second samples.

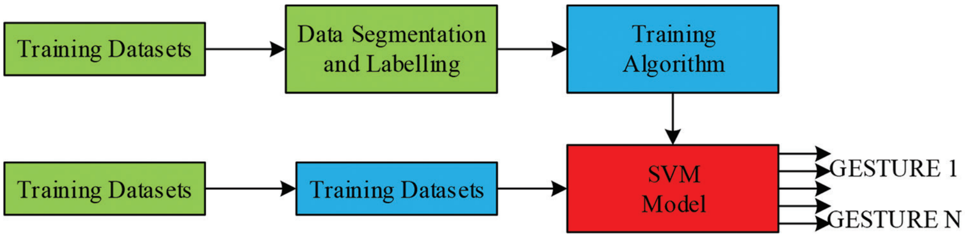

Fig. 5. Block diagram of the support vector machine system

4.2. Killer Whale Optimization Algorithm

Fig. 6 depicts the fitness graph for the Killer Whale Optimization algorithm; the fitness value at iteration 1 is 0.0920, showing that the algorithm is in the early stages of investigating possible solutions. By iteration 2, the fitness value has improved to 0.0751, indicating that the algorithm is effectively refining its search and discovering superior solutions. From iterations 3 to 5, the fitness value remains constant at 0.0751, indicating that the algorithm has hit a local optimum and is performing consistently without major improvements. Between iterations 6 and 10, the fitness value drops significantly to 0.0750, indicating that the algorithm is still engaged in exploration and exploitation, but the benefits are minor at this time.

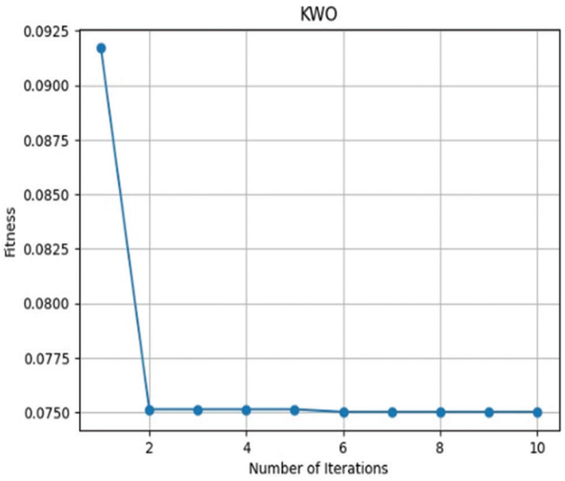

Fig. 6. Killer Whale Optimization (KWO) algorithm

4.3. Confusion Matrix

Fig. 7 compares a model's predictions to real emotions in a dataset, indicating both strengths and opportunities for development. The model guessed "none" 210 times, "angry" 35 times, "happy" 146 times, "calm" 66 times, and "fearful" 234 times. Diagonal values indicate good predictions, and higher values imply better classification accuracy. For example, the model accurately predicted "fearful" 234 times, indicating excellent performance. Off-diagonal values show misclassifications and provide areas for improvement. Overall, the model's performance is evaluated to identify its strengths and opportunities for development.

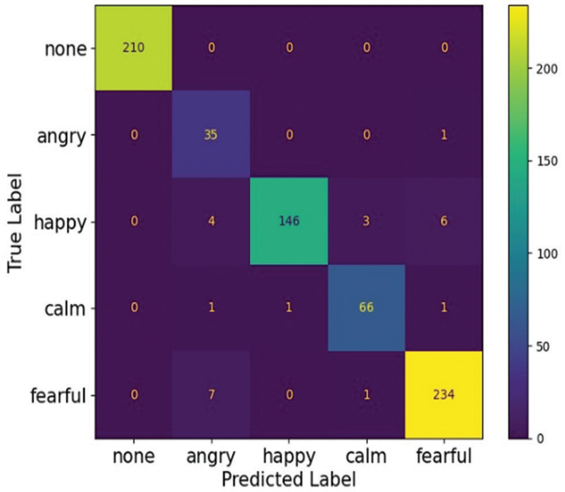

Fig. 7. Graph of the confusion matrix

4.4. Receiver Operating Characteristic Curve Under the Area Under the Curve

The true positive rate (TPR), also known as sensitivity or recall, quantifies the proportion of actual positives properly detected by the model. It is calculated using the following formula in Equation (26):

$$TPR = \frac{True\,Positive\;(TP)}{True\,Positive\big(TP\big) + False\,Negative\;(FN)} \quad (26)$$

The false-positive rate (FPR) is the percentage of actual negatives that are wrongly recognized as positives by the model. It is calculated using the following formula in Equation (27):

$$FPR = \frac{False\,Positive\;(TP)}{False\,Positive\big(TP\big) + True\,Negative\;(FN)} \quad (27)$$

Fig. 8 depicts the relationship between the TPR and the FPR, with each axis ranging from 0.0 to 1.0. The area under the curve (AUC) for each class is shown below: Class Zero has an AUC of 100%, Class One has an AUC of 99%, Class Two has an AUC of 98%,

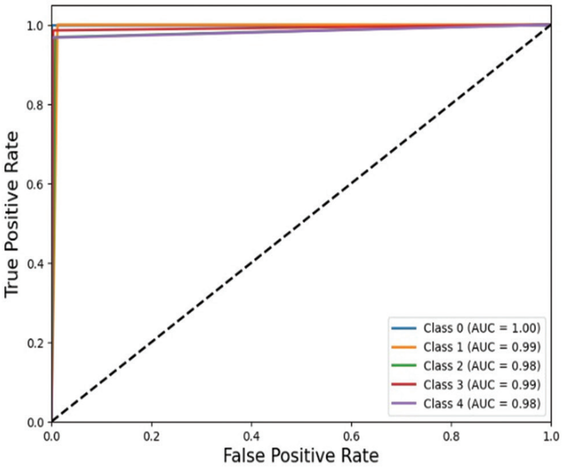

Fig. 8. Receiver operating characteristic curve under the area under the curve

Class Three has an AUC of 99%, and Class Four has an AUC of 98%. These AUC values indicate the classifier's performance, with higher values suggesting a stronger ability to differentiate between classes.

4.5. Performance Metrics

The performance metrics include a variety of critical indicators for assessing a model's success.

Fig. 9 displays the performance metrics, with the x-axis representing the score ranging from 0.0 to 1.0. The results achieved are as follows: accuracy 98%, precision 98%, recall 98%, F1-score 98%, sensitivity 99%, and specificity 99%.

4.6. Precision for Each Class

The precision formula calculates the accuracy of a model's positive predictions. The definition goes as follows in Equation (28):

$$Precision = \frac{TP}{TP + FP} \tag{28}$$

Fig. 10 shows the precision reached for each emotion class, with x-axis values ranging from 0.0 to 1.0. The emotions are divided into five categories: angry, none, happy, calm, and fearful. The precision for each class is as follows: angry 99%, none 82%, happy 98%, calm 98%, and fearful 99%.

4.7. F1-Score for Each Class

The F1-score combines precision and recall into a single metric. It is very useful when trying to achieve a balance between precision and recall, particularly if your class distribution is asymmetrical (Equation [29]).

$$F1\,Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{29}$$

Fig. 11 shows the F1-score reached for each emotion class, with x-axis values ranging from 0.0 to 1.0. The emotions are divided into five categories: angry, none, happy, calm, and fearful. The F1-score for each class is as follows: angry 99%, none 90%, happy 97%, calm 98%, and fearful 98%.

4.8. Recall for Each Class

Recall, also known as sensitivity or true positive rate, is a metric that measures a model's ability to accurately identify all relevant instances in a dataset. It is highly useful for reducing false negatives (Equation [30]).

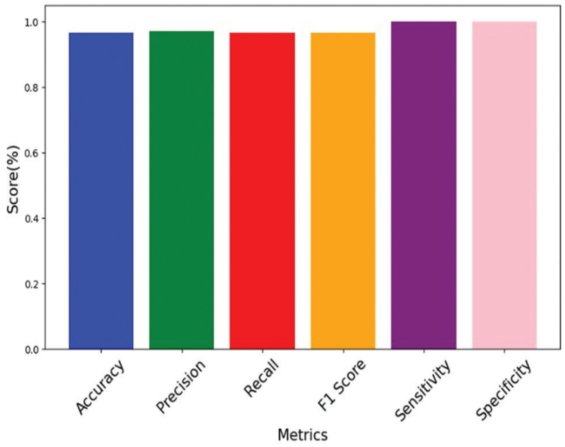

Fig. 9. Performance metrics

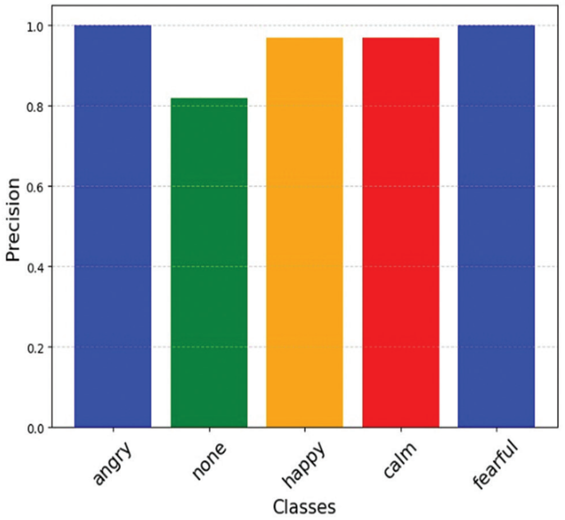

Fig. 10. Precision for each class

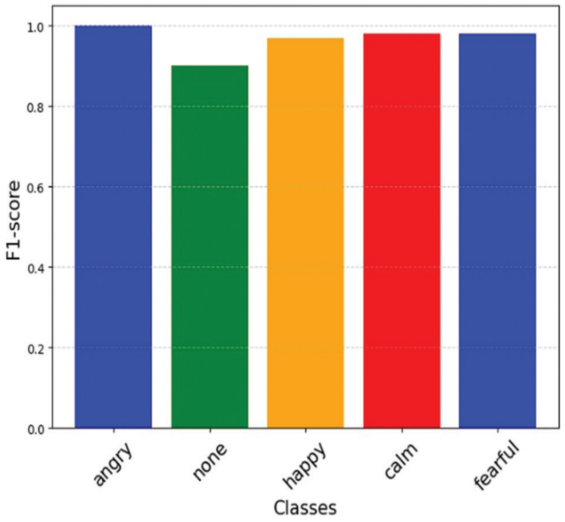

Fig. 11. F1-score for each class

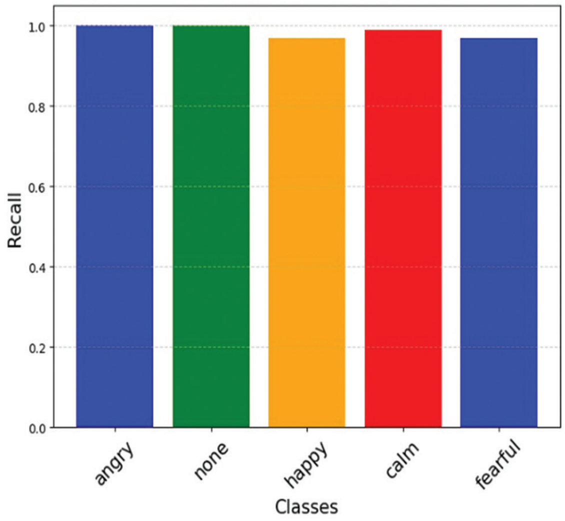

Fig. 12. Recall for each class

$$Recall = \frac{TP}{TP + FN} \tag{30}$$

It is vital in situations when missing a positive instance is more expensive or significant than mistakenly identifying a non-positive example.

Fig. 12 depicts the recall scores for each emotion class, with x-axis values ranging from 0.0 to 1.0. The emotions are classified into five types: angry, none, happy, calm, and fearful. The recall for each class is as follows: angry 99%, none 99%, happy 97%, calm 98%, and fearful 96%.

4.9. Comparative Analysis

This section demonstrates that the suggested methodology outperforms alternative models with fewer parameters, such as logistic regression

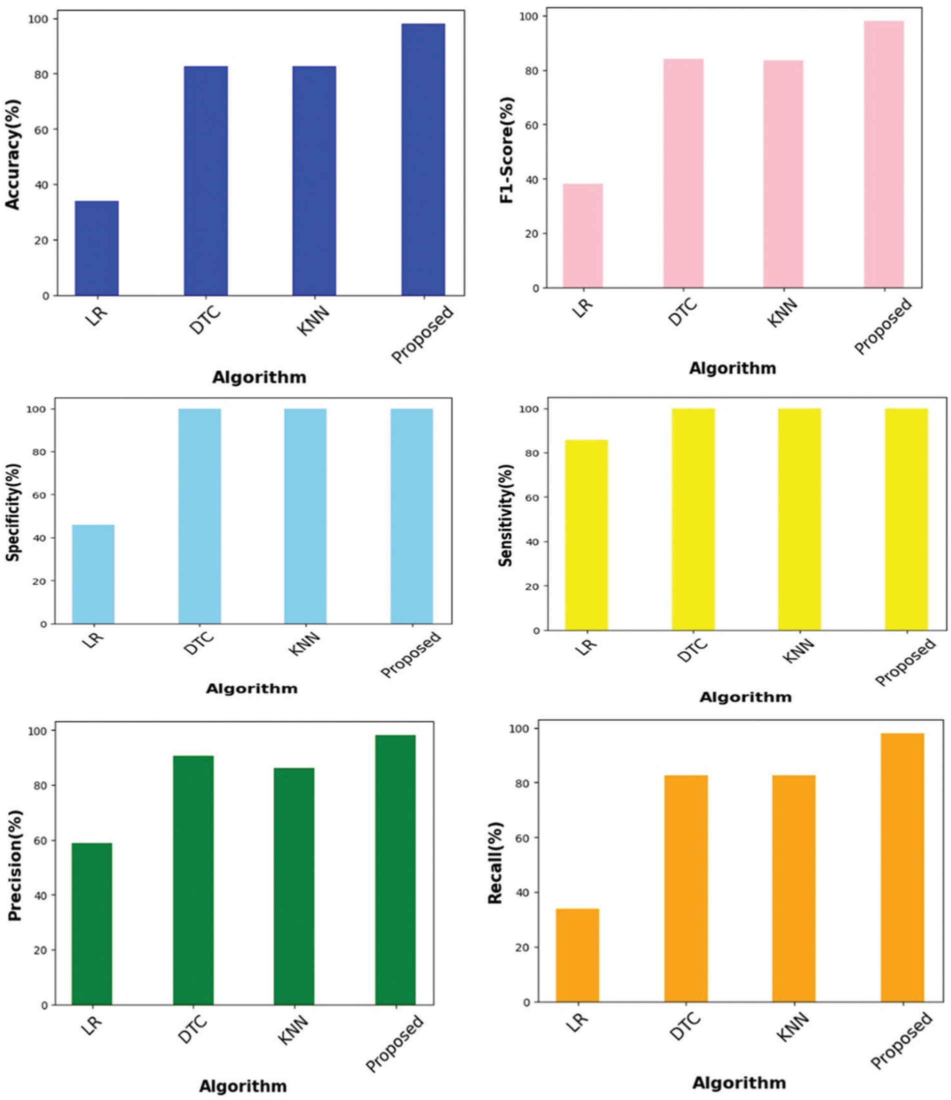

Fig. 13. Comparative analysis
Abbreviations: DTC: Decision tree classifier; KNN: K-nearest neighbor; LR: Logistic regression

Table 1. Comparative analysis

Methods	Accuracy%	F1-score%	Specificity%	Sensitivity%	Precision%	Recall%
Logistic regression	35	38	45	85	60	35
Decision tree classifier	82	84	99	99	90	83
K-nearest neighbors	80	82	99	99	85	83
Proposed	98	98	99	99	98	98

(Luna-Jiménez et al., 2021), decision tree classifier (Amartya & Kumar, 2022), and K-nearest neighbors (Subbarao et al., 2021).

Table 1 and Fig. 13 compare the performance of Marathi SER algorithms in terms of accuracy, F1-score, specificity, sensitivity, precision, and recall, demonstrating their usefulness. The logistic regression model achieved 35% accuracy, 38% F1-score, 45% specificity, 85% sensitivity, 60% precision, and 35% recall. The decision tree classifier model performed better, with an accuracy of 82%, an F1-score of 84%, a specificity of 99%, a sensitivity of 99%, a precision of 90%, and an 83% recall. The K-nearest neighbors approach achieved 80% accuracy, 82% F1-score, 99% specificity, 99% sensitivity, 85% precision, and 83% recall. In contrast, the proposed technique surpassed all other models, with 98% accuracy, 98% F1-score, 99% specificity, 99% sensitivity, 98% precision, and 98% recall.

5. Conclusion

The study's objectives are effectively addressed by the suggested methodology for SER in Marathi, which overcomes the shortcomings of current approaches and captures the distinctive grammatical and emotional subtleties of Marathi speech. The model attains remarkable performance metrics, including 98% accuracy, 98% F1-score, 99% specificity, 99% sensitivity, 98% precision, and 98% recall, using sophisticated signal processing, thorough feature extraction, and a novel classification strategy that combines an attentive DBN with an SVM. The study's shortcomings offer the potential for further research, even if the methodology addresses important issues and establishes a new standard for SER in Marathi. These include adding more languages, improving real-time processing for mobile apps, strengthening resilience to various noise and acoustic conditions, incorporating multimodal data such as physiological signals and facial expressions, and utilizing innovative architectures like transformers to further improve performance.

Acknowledgments

The authors would like to thank the Deanship of RMD Sinhgad School of Engineering for supporting this work.

References

Abdel-Hamid, L., Shaker, N.H., Emara, I. (2020). Analysis of linguistic and prosodic features of bilingual Arabic-English speakers for speech emotion recognition. *IEEE Access*, 8, 72957–72970.

Abdusalomov, A., Kutlimuratov, A., Nasimov, R., & Whangbo, T.K. (2023). Improved speech emotion recognition focusing on high-level data representations and swift feature extraction calculation. *Computers, Materials and Continua*, 77(3), 2915–2933.

Akçay, M.B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56–76.

Akinpelu, S., & Viriri, S. (2024). Deep learning framework for speech emotion classification: A survey of the state-of-the-art. *IEEE Access*, 12, 152152.

Alam Monisha, S.T., & Sultana, S. (2022). A review of the advancement in speech emotion recognition for indo-aryan and dravidian languages. In: *Advances in Human-Computer Interaction*. Wiley, Hoboken.

Alluhaidan, A.S., Saidani, O., Jahangir, R., Nauman, M.A., & Neffati, O.S. (2023). Speech emotion recognition through hybrid features and convolutional neural network. *Applied Sciences*, 13(8), 4750.

Amartya, J.G.M., & Kumar, S.M. (2022). Speech emotion recognition in machine learning to improve accuracy using novel support vector machine and compared with decision tree algorithm. *Journal of Pharmaceutical Negative Results,* 185–192.

Arul, V.H. (2021). Deep learning methods for data classification. In: *Artificial Intelligence in Data Mining*. Academic Press, p87–108

Bachate, R.P., Sharma, A., Singh, A., Aly, A.A., Alghtani, A.H., & Le, D.N. (2022). Enhanced marathi speech recognition facilitated by grasshopper optimisation-based recurrent neural network. *Computer Systems Science and Engineering,* 43(2), 439–454.

Bhangale, K., & Kothandaraman, M. (2023). Speech emotion recognition based on multiple acoustic features and deep convolutional neural network. *Electronics*, 12(4), 839.

Byun, S.W., & Lee, S.P. (2021). A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Applied Sciences*, 11(4), 1890.

Chai, J., Zeng, H., Li, A., & Ngai, E.W. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6, 100134.

Chaudhari, P., Nandeshwar, P., Bansal, S., & Kumar, N. (2023). MahaEmoSen: Towards Emotion-aware Multimodal Marathi Sentiment Analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(9), 1–24.

Er, M.B. (2020). A novel approach for classification of speech emotions based on deep and acoustic features. *IEEE Access*, 8, 221640–221653.

Farooq, M., Hussain, F., Baloch, N.K., Raja, F.R., Yu, H., & Zikria, Y.B. (2020). Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. *Sensors*, 20, 6008.

Hammed, F.A., & George, L. (2023). Using speech signal for emotion recognition using hybrid features with SVM classifier. *Wasit Journal of Computer and Mathematics Science*, 2(1), 27–38.

Harhare, T., & Shah, M. (2021). Linear mixed effect modelling for analyzing prosodic parameters for marathi language emotions. *International Journal of Advanced Computer Science and Applications*, 12(12).

Kaur, K., & Singh, P. (2023). Comparison of various feature selection algorithms in speech emotion recognition. *AIUB Journal of Science and Engineering (AJSE)*, 22(2), 125–131.

Kawade, R., & Jagtap, S. (2024). Indian cross corpus speech emotion recognition using multiple spectral-temporal-voice quality acoustic features and deep convolution neural network. *Revue d'Intelligence Artificielle*, 38(3), 913–927.

Kishor, B., Mohanaprasad, K. (2022). Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network. In: *Futuristic Communication and Network Technologies*. Springer, Singapore, p241–250.

Kok, C.L., Ho, C.K., Tan, F.K., & Koh, Y.Y. (2024). Machine learning-based feature extraction and classification of emg signals for intuitive prosthetic control. *Applied Sciences*, 14(13), 5784.

Li, R., Zhao, J., & Jin, Q. (2021). Speech Emotion Recognition Via Multi-Level Cross-Modal Distillation. In: *Proceedings of Interspeech*, p4488–4492.

Li, Z., Huang, H., Zhang, Z., & Shi, G. (2022). Manifold-based multi-deep belief network for feature extraction of hyperspectral image. *Remote Sensing*, 14(6), 1484.

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10(10), 1163.

Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J.M., & Fernández-Martínez, F. (2021). A proposal for multimodal emotion recognition using aural transformers and action units on raves dataset. *Applied Sciences*, 12(1), 327.

Madanian, S., Chen, T., Adeleye, O., Templeton, J.M., Poellabauer, C., Parry, D., & Schneider, S.L. (2023). Speech emotion recognition using machine learning-a systematic review. *Intelligent Systems with Applications*, 20, 200266.

Oh, S., & Kim, D.K. (2022). Comparative analysis of emotion classification based on facial expression and physiological signals using deep learning. *Applied Sciences*, 12(3), 1286.

Padman, S., & Magare, D. (2022). Regional language speech emotion detection using deep neural network. *ITM Web of Conferences*, 44, 03071.

Papala, G., Ransing, A., & Jain, P. (2023). Sentiment analysis and speaker diarization in hindi and marathi using finetuned whisper: Sentiment analysis in Hindi and Marathi. *Scalable Computing: Practice and Experience*, 24(4), 835–846.

Sajjad, M., & Kwon, S. (2020). Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. *IEEE Access*, 8, 79861–79875.

Shah, F.M., Ranjan, A., Yadav, J., Deepak, A. (2021). A survey of speech emotion recognition in the natural environment. *Digital Signal Process*, 110, 102951.

Singh, Y.B., & Goel, S. (2021). An efficient algorithm for recognition of emotions from speaker and language independent speech using deep learning. *Multimedia Tools and Applications*, 80(9), 14001–14018.

Sonawane, S., & Kulkarni, N. (2020). Speech emotion recognition based on MFCC and convolutional neural network. *International Journal of Advance Scientific Research and Engineering Trends*, 5, 18–22.

Subbarao, M.V., Terlapu, S.K., Geethika, N., & Harika, K.D. (2021). Speech emotion recognition using k-nearest neighbor classifiers. *In: Recent Advances in Artificial Intelligence and Data Engineering: Select Proceedings of AIDE*.

Springer Verlag, Singapore, p123–131.

Tiwari, P., Dehdashti, S., Obeid, A.K., Marttinen, P., & Bruza, P. (2022). Kernel method based on non-linear coherent states in quantum feature space. *Journal of Physics A: Mathematical and Theoretical*, 55(35), 355301.

Yang, Z., Zhou, S., Zhang, L., & Serikawa, S. (2024). Optimizing Speech Emotion Recognition with Hilbert Curve and convolutional neural network. *Cognitive Robotics*, 4, 30–41.

Zaidi, S.A.M., Latif, S., & Qadi, J. (2023). *Cross-Language Speech Emotion Recognition Using Multimodal Dual Attention Transformers*. [arXiv Preprint].

AUTHOR BIOGRAPHIES

Varsha Gaikwad completed her Bachelor's Degree in Electronics and Telecommunication Engineering from BMIT, Solapur University in 2010 and Master's Degree in Electronics and Telecommunication Engineering with specialization in Signal Processing from SPPU Pune University in 2014. At present, she is working as an Assistant Professor in STES's RMD Sinhgad School of Engineering, Pune. She is a Life Member of ISTE, IAEG India. Her areas of interest are Speech Signal Processing and Digital Communication, Machine and Deep Learning, and Pattern Recognition. She has attended many national and international conferences. She has published six papers in the International Journal, two papers in International Conferences, and two papers in National Conferences. She has 6-year Professional Experience as Assistant Professor and she has 4-year Professional Experience as Lecturer.