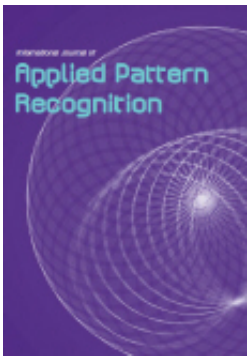


International Journal of Applied Pattern Recognition

ISSN online: 2049-8888 - ISSN print: 2049-887X
https://www.inderscience.com/ijapr

The speech emotion recognition in multi-languages using an ensemble deep learning-based technique

Putta Aruna Kumari, Balusu Nandini

DOI: 10.1504/IJAPR.2024.10071634

Article History:

Received:	09 December 2024
Last revised:	26 March 2025
Accepted:	30 April 2025
Published online:	19 June 2025

The speech emotion recognition in multi-languages using an ensemble deep learning-based technique

Putta Aruna Kumari* and Balusu Nandini

Department of Computer Science,
Telangana Social Welfare Residential Institution for Women,
Nizamabad, Telangana State, 503001, India
Email: aruna.skurapati@gmail.com
Email: nandinibalusu@telanganauniversity.ac.in
*Corresponding author

Abstract: Accurate emotion detection from speech signals is essential for enhancing human-computer interaction (HCI) systems. However, existing SER methods often suffer from poor feature representation and limited dataset diversity, resulting in suboptimal performance. To address these challenges, this paper proposes an advanced deep bottleneck residual convolutional neural network (DBR-CNN) integrated with the SEResNeXt-101 feature extraction framework and optimised using the coati optimisation algorithm (COA). The model is trained and evaluated on four diverse benchmark datasets: URDU, EMO-DB, EMOVO, and SAVEE. In the pre-processing phase, speech signals undergo noise reduction and normalisation to enhance data quality. The SEResNeXt-101 extractor then captures high-level features with reduced complexity, which are subsequently processed by the DBR-CNN to classify emotions with greater accuracy. The COA fine-tunes the model to improve classification efficiency. Experimental results demonstrate that the proposed model significantly outperforms state-of-the-art (SOTA) methods. The optimised DBR-CNN framework provides robust, scalable, and highly accurate emotion recognition.

Keywords: speech emotion recognition; SER; feature extraction; deep bottleneck residual convolutional neural network; DBR-CNN; noise reduction; coati optimisation algorithm; COA; speech emotion datasets; SEResNeXt-101; deep learning.

Reference to this paper should be made as follows: Kumari, P.A. and Nandini, B. (2024) 'The speech emotion recognition in multi-languages using an ensemble deep learning-based technique', *Int. J. Applied Pattern Recognition*, Vol. 7, Nos. 3/4, pp.263–295.

Biographical notes: Putta Aruna Kumari is currently working as degree Lecturer in the Department of Computer Science, Telangana Social Welfare Residential Institution for Women, Nizamabad, Telangana State, India. She completed BSc (Computer Science) and MCA from Osmania University, Hyderabad. Her research work mainly focuses on machine learning, deep learning and artificial intelligence.

Balusu Nandini has completed her PhD from Jawaharlal Nehru Technological University, Hyderabad. Her current research focuses on artificial intelligence, block chain technologies, machine learning, deep learning and computer networks. Currently, she is working as degree Lecturer in the Department of

Computer Science, Telangana Social Welfare Residential Institution for Women, Nizamabad, Telangana State, India. She has published papers in SCI, Scopus, UGC Care listed national, international journals and conferences.

1 Introduction

Speech emotion recognition (SER) is a rapidly developing field of study in affective computing that aims to extract an emotional state from a person's speech to understand the semantics of an uttered sentence. SER has many possible uses, including a call centre service (Ancilin and Milton, 2021; Singh et al., 2021). The system can react respectfully to customers once it has identified their emotions from their tone of voice. On the other hand, a car-board system can recognise stress indicators in a driver's voice and alert them for careful driving. Moreover, SER can give feedback on activities and interactions between people and robots (Li et al., 2021; Zhang et al., 2021; Zehra et al., 2021).

Current psychological research has shown how people can still identify emotions in cross-linguistic conversations even when they cannot understand the language (Krishnan et al., 2021). Interest in SER has recently shifted from monolingual to multilingual settings, as might be expected given the current SER systems' excellent generalisation skills and accuracy gains. The goal is to differentiate speech emotions throughout languages simply for listeners to comprehend and react to. The development of emotional systems in real-world contexts could be significantly accelerated by this focus (Kwon, 2021; Gerczuk et al., 2021; Abdulmohsin et al., 2021; Yildirim et al., 2021).

There have been numerous attempts to identify emotions from speech, with significant results in terms of identification. Here are some further examples of relevant research (Chen et al., 2023; Liu et al., 2023). However, it was found that depending on the language, different voice characteristics perform well as SER in specific compositions. This scenario demonstrates how challenging it is to transform current monolingual SER methods into multilingual SER tasks, as changing the source language requires retraining a system and reselecting a set of optimal acoustic features (Bhangale and Kothandaraman, 2023). Recently, a large amount of research has been developed to address the issues of linguistic variations. The authors focused on establishing 'similar' situations across joint corpora by transferring adaption mechanisms. It was demonstrated that while neutral voice positions and other emotional states have similar positions and divisions across languages, neutral speech positions vary (Shahin et al., 2023; Sun et al., 2023). With this knowledge, finding the shared perceptual characteristics between different languages will significantly improve the task of SER in the multilingual context.

Despite advancements, current SER systems still face notable challenges. While early models focused primarily on monolingual datasets, modern applications demand robust systems capable of handling multilingual and cross-linguistic scenarios. This demand arises from the observation that human perception of emotion in speech transcends linguistic barriers, prompting research into generalised models that can accurately classify emotions across diverse languages and acoustic environments. However, existing SER frameworks often struggle with generalisation due to limitations in feature extraction, inadequate data diversity, and model overfitting when applied across varied datasets.

Motivated by these challenges, this study introduces a novel approach that combines an advanced feature extraction framework (SEResNeXt-101), a deep bottleneck residual convolutional neural network (DBR-CNN), and an optimisation step using the coati optimisation algorithm (COA) to improve SER performance across multiple datasets and linguistic contexts. The proposed model is designed to address the limitations of existing techniques by enhancing the ability to capture and process key temporal and spectral information within speech signals, ultimately improving classification accuracy and model generalisability.

1.1 Research motivation

The motivation behind this research stems from the growing need for highly accurate and generalisable SER systems capable of functioning effectively across diverse linguistic and acoustic environments. While existing SER models have demonstrated promising results, they often suffer from poor cross-lingual adaptability, limited feature representation, and inconsistent performance across different datasets. The ability to automatically recognise human emotions from speech is crucial for enhancing human-computer interaction (HCI) in real-world applications such as intelligent virtual assistants, call centres, automotive safety systems, and social robotics. However, most traditional models lack the robustness to handle variations in speech signals caused by language, speaker diversity, and environmental noise. To address these gaps, this study is motivated to design a deep learning-based SER framework that not only improves the accuracy of emotion classification but also enhances the model's capacity to generalise across multilingual datasets. By integrating an advanced feature extractor (SEResNeXt-101), a deep bottleneck residual CNN (DBR-CNN), and an optimisation algorithm (COA), the research aims to deliver a highly efficient and scalable solution that meets the demands of modern SER applications.

1.2 Novelty of the proposed model

- Collect the speech data from four different datasets. Then, noise reduction and normalisation will be performed in the pre-processing stage.
- To enhance the accuracy and efficiency of recognising emotions in speech, the proposed approach involves utilising the SEResNeXt-101 technique to extract features from the audio data.
- Then, the attributes retrieved from the speech signals are utilised as inputs to the DBR-CNN to recognise various emotional categories. The DBR-CNN model employs its architecture, including bottleneck layers and residual connections, to process the extracted features.
- The proposed approach integrates the COA to enhance recognition and classification performance. This algorithm, renowned for its efficiency in optimising complex systems, aids in refining the model's parameters and decision boundaries.

1.3 Organisation of the paper

Section 2 of this paper provides an overview of the relevant SER works on speech and text modalities. Section 3 delves into the specifics of the deep learning model that is being presented – finally, Section 4 and Section 5 show the results and conclusions of the experiments.

2 Literature survey

According to many research articles, speech is the most researched medium for emotion recognition (16–20). Ahmed, et al., 2023 suggested an ensemble that employed the combined predictive performance of three distinct structures, the LSTM, CNN, and GRU, for efficient extraction of features. First, 1D CNN was used in the model, and then fully connected networks (FCN) were used. The GRU-FCN and LSTM-FCN layers follow the CNN layer in the other two structures. They have extended the data to improve model generalisation by adding pitch shifting, additive white Gaussian noise, and signal level stretching. Five kinds of attributes were retrieved from the speech samples: every audio file in those datasets had the following five characteristics: log mel-scaled spectrogram, mel-frequency cepstral coefficients, root mean square value, zero-crossing rate, and chromatogram. Extracting local and long-term global contextual depictions of speech signals is the primary goal of the three models.

Xu et al. (2022) suggested a hybrid deep learning with a multi-type features model (HD-MFM) to integrate speech temporal, acoustic, and picture data. They specifically used CNN to retrieve picture data from speech spectrograms. The acoustic data were extracted from the statistical aspects of speech using a deep neural network (DNN). Next, temporal data was extracted from the speech MFCC using LSTM. Concatenating three distinct speech feature types yields a richer emotion description with improved discriminative properties. Then, they examine two fusion strategies – separating and merging – because they impact the relationship between characteristics.

The concurrent spatial-temporal and grammatical (CoSTGA) model, which Kakuba et al. (2022) suggested a DL-based model that simultaneously learns temporal, spatial, and semantic representations in the LFLB and fused as a latent vector to create an input to the GFLB. Utilising the MLTED, which they previously made, they evaluate the effectiveness of multi-level feature fusion compared to single-level fusion. The suggested CoSTGA model combines multi-level fusion two times: once at the GFLB level, where the spatial-temporal characteristics are fused with the semantic tendency attributes, and once at the LFLB level, where similar attributes are retrieved separately from a modality. DCC, multi-head, BiLSTM, TE, and self-attention mechanisms are all included in the suggested CoSTGA framework. Andayani et al. (2022) suggested a hybrid transformer encoder and LSTM network to identify long-term relationships in speech signals and categorise emotions. The suggested hybrid LSTM-Transformer classifier receives speech attributes derived using the mel frequency cepstral coefficient (MFCC). The suggested LSTM-Transformer model was subjected to some performance assessments.

A technique to increase the accuracy of classified eight emotions from the human voice was suggested by Jothimani et al. (2022). The suggested MFF-SAug research uses white noise injection, pitch tuning, and noise removal to improve speech emotion forecasting. ZCR, MFCC, and RMS are three feature extraction approaches applied and

integrated into pre-processed speech signals to obtain significant performance for emotion identification. For improved voice emotion recognition and speech representation learning, a convolution neural network (CNN) was suggested.

The summary of the literature survey mentioned in Table 1.

Table 1 Summary of the literature survey

Reference	*Methods*	*Merits*	*Demerits*
Ahmed et al. (2023)	CNN-LSTM-GRU	Combines the strengths of CNN, LSTM, and GRU for better feature extraction; employs data augmentation for improved generalisation; extracts both local and global contextual features	High computational complexity due to multiple deep learning architectures; risk of overfitting despite augmentation
Xu et al. (2022)	HD-MFM	Integrates temporal, acoustic, and visual features using CNN, DNN, and LSTM; richer emotion representation via multi-feature fusion; improved discriminative power by combining MFCC, spectrogram, and statistical speech features	Complex fusion process requiring fine-tuning of multiple networks; increased training time and computational resources
Kakuba et al. (2022)	CoSTGA	Simultaneous learning of temporal, spatial, and semantic features; multi-level fusion at both early and late stages; includes attention mechanisms and BiLSTM for enhanced context awareness	Complex model architecture with several fusion layers increases model size and inference time; challenging to implement in real-time systems
Andayani et al. (2022)	LSTM-transformer hybrid model	Effectively captures long-term dependencies in speech signals; leverages MFCC features; combines strengths of LSTM and Transformer architectures to enhance sequential and contextual learning	Transformer integration increases model complexity and resource requirements; may not perform well with limited or imbalanced datasets
Jothimani et al. (2022)	CNN with MFF-SAug	Improves robustness using speech augmentations like noise injection and pitch tuning; applies ZCR, MFCC, and RMS for multi-feature extraction; CNN ensures effective feature learning and representation	Focuses primarily on low-level features; augmentation methods may not generalise well across highly varied real-world datasets

2.1 Problem statement

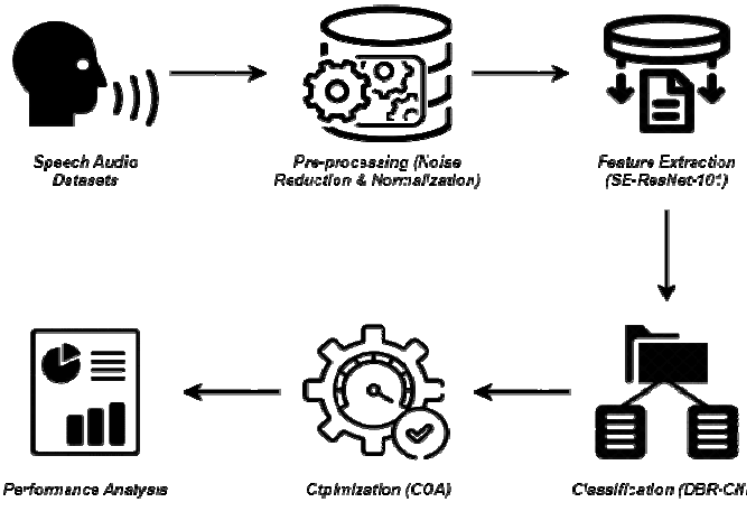
Previously, emotions have been categorised by either utilising a single machine learning technique or extracting many characteristics, resulting in long computing times. This has limited us to relying on a single classifier that has demonstrated lower accuracy than anticipated, preventing us from utilising the data every classifier has to offer. An optimal approach for extracting relevant and discriminative features from multilingual SER is absent. It takes a while to set up in a peaceful environment and gather enormous volumes

of labelled data for local languages. Additional challenges facing SER systems originate from the classifiers' incapacity to be broadly applied. Some classifiers often show a rapid loss in classification capacity when tested with speech data obtained from various speakers, linguistic content, domain contexts, and acoustic conditions. The diversity between speakers and the variety of phrases affect SER systems' performance. We present our proposed methodology for handling the problem above.

3 Proposed methodology

The comprehensive study approach used to identify emotions from audio data is covered in this section. Many real-world applications, including computer games, mobile services, HCI, and emotion assessment, have used data science in recent years. SER is a novel and challenging research area among the many applications. According to recent studies, handcrafted features for SER yield the highest results, but when applied in complicated circumstances, they fail to give accuracy. The SER that extracts attributes from voice signals was developed using a deep learning method. Although deep learning-based SER techniques resolve accuracy problems, the published methods still have a lot of holes. A unique SER model was presented to address the constraints previously discussed in this study.

Figure 1 An architecture of proposed methodology



The first step involves collecting data from speech samples labelled with corresponding emotions from four datasets named URDU, Berlin Emo − DB, EMOVO, and SAVEE. These datasets should cover various speakers, emotions, and environmental conditions to ensure the model's robustness. Then, normalisation and noise reduction in speech audios are performed in the pre-processing stage. The extraction of features can reduce computational error, computational time, and model complexity in voice emotion categorisation. Nonetheless, the SEResNeXt-101 technique must be used to extract features that offer reliable information about the mood. Then, the extracted features from the speech are used to recognise the emotional categories using DBR-CNN. Finally, the

COA is employed to improve the recognition and classification performance. Figure 1 shows the architectural structure of the proposed model.

3.1 Pre-processing

In a SER system, pre-processing comes right after data collection and is utilised to train the classifier. While some pre-processing methods extract attributes, others normalise the data to ensure that differences in speakers and recordings won't interfere with the identification process.

3.1.1 Noise reduction

In the real world, both the speech signal and the surrounding noise are recorded. This impacts the identification rate, so it's necessary to use some noise reduction methods to eliminate or lessen the noise. The clean signal is estimated from the noisy signal's sample function. It requires apriori speech and noise spectrum data. It is predicated on the idea that estimates of the speech and additive noise spectrum are known. The technique minimises the predicted distortion between the estimated and clean voice signals.

3.1.2 Normalisation

Reducing speaker and recording variability without sacrificing the attributes' capacity to discriminate is accomplished in part by data normalisation. The generalisation capacity of features is improved by feature normalisation. Normalisation can occur at several levels, including the corpus and function. Z-normalisation (standard score) is the most popular normalisation technique. When the data's mean (μ) and standard deviation (σ) are known, z-normalisation can be computed as $z = x–\mu/\sigma$.

Figure 2 shows the outcome of the pre-processing stage of the original audio spectrum and noise-removed audio spectrum.

3.2 Feature extraction

SER depends primarily on attributes. The recognition rate is enhanced by a well-designed set of attributes that effectively capture every emotion. Despite several features being employed for SER systems, there has yet to be a widely recognised set of attributes for accurate and unique categorisation. In SER, there are various advantages to use SEResNeXt-101 for feature extraction. Adaptation to emotional content is made more accessible by its transfer learning capabilities, which use pre-trained models for generic feature extraction. Its hierarchical representation learning allows the capture of complicated patterns in voice signals. Compared to conventional techniques or shallower architectures, SEResNeXt-101 may improve emotion categorisation due to its ability to acquire contextual cues and state-of-the-art (SOTA) performance in visual recognition challenges. It is also a potential method for reliable and accurate emotion recognition of speech systems because of its flexibility, which enables scalability and change based on task complexity and available resources.

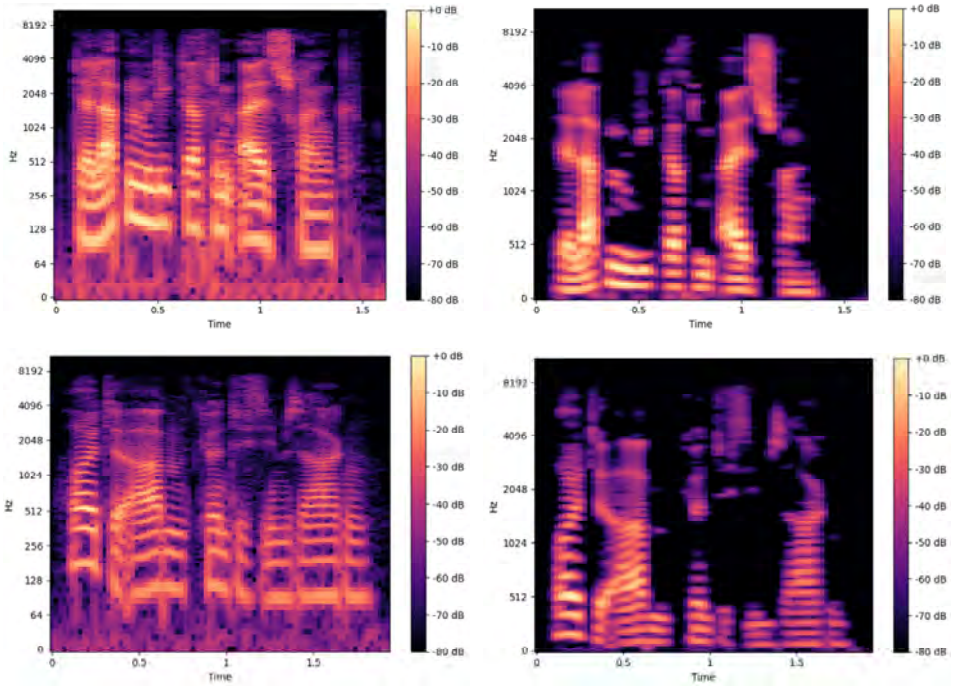
The input audio data was processed using a pre-trained SE-ResNeXt-101 model to acquire attributes for speech emotion identification. A squeeze-and-excite module was added to the ResNeXt101-32x4d variation, which became the SE-ResNeXt-101-32x4d.

The computational block used to create the squeeze-and-excitation block was E_{ks}, which transforms the input $Z \in S^{B' \times Y' \times R'}$ to attribute mapping $V \in S^{B \times Y \times R}$. The learned sets of filter kernels were represented as $U = [u_1, u_2, \ldots, u_R]$ in the following characters, where u_R stood for the parameter of the R^{th} filter. E_{ks} was a convolutional operator. Equation (1) can, therefore, be used to describe the output as $V = [v_1, v_2, \ldots, v_R]$:

$$v_r = u_r * Z = \sum_{m-1}^{R'} u_r^m * Z^m \tag{1}$$

Convolution is represented by *, Z is represented by $[z_1, z_2, \ldots, z^{R'}]$, and the 2D spatial kernel is defined by $vr \in S^{B \times Y} \cdot u_r^m$, which means a single u_r channel that operates on the corresponding Z channels.

Figure 2 Pre-processing on audio emotion file (a) original audio spectrum (b) denoised audio spectrum (see online version for colours)



(a) (b)

To make the notation simpler, bias terms were removed. Channel dependencies were unintentionally stored in u_r because the result is the sum of all the channels. One potential solution to the problem of channel dependency exploitation to compress global geographical data into channel descriptors. Global average pooling was used to generate channel-specific data to achieve this. Formally, by reducing V across its spatial dimensions $B \times Y$, the statistic $a \in S^R$ was generated, and equation (2) yielded the r^{th} element of a.

$$a_r = E_{md}(v_r) = \frac{1}{B \times Y} \sum_{j-1}^{B} \sum_{t-1}^{Y} v_r(j, i) \tag{2}$$

where v_r represents the feature map for the r^{th} channel, a_r the r^{th} element of the generated statistic, a, which represents the global descriptor for channel-specific information, $B \times Y$ the spatial dimensions of the feature map, E_{md} the expectation (mean) operation applied across spatial dimensions.

A second operation was carried out to capture channel-wise dependencies using the data collected during the squeeze operations to capture channel-wise dependencies fully. An adaptation of ResNet that bore a striking resemblance to the Inception model was the ResNeXt module. Both follow the split-transform-merge approach, except that in this version, the output of distinct routes was combined rather than depth-concatenated as in the inception model. Experiments showed that increasing cardinality was more accurate than expanding or deepening the search.

$$e(z) = \sum_{j-1}^{R} K_j(z) \tag{3}$$

$K_j(z)$ represented function in equation (3). Projecting z into a (perhaps low dimensional) space, K_j modifies and embeds it, resembling a primary neuron. R indicates how big the set of changes to be combined is in equation (3). R is referred to as having cardinalities. The dimension of cardinality determines how many more complex transformations there are. Equation (4)'s aggregated transformation represents the residual function.

$$f = z + \sum_{j-1}^{R} K_f(z) \tag{4}$$

This represents an aggregated transformation where the input z is modified by multiple, $K_f(z)$ functions and summed to form the final residual function. The residual connection ensures that the original input, z is retained while adding transformed variations through the summation term.

3.3 Classification

A DBR-CNN is employed to recognise speech emotion in audio files. Firstly, DBR-CNNs enable more complex and abstract attributes to be learned efficiently by training deeper networks with residual connections. This may improve the model's capacity to identify small differences in speech patterns associated with various emotional states. Second, by compressing feature representations while maintaining crucial information, the bottleneck layers in DBR-CNNs help lower computational costs and memory requirements. A bottleneck residual CNN structure is proposed that combines average pooling, convolutional, and fully connected hidden layers in its 76 hidden layers, three parallel residual blocks, and a unique design. A specific architectural component commonly used in DNNs, particularly CNNs, is called a 'bottleneck' or 'bottleneck layer' in layered models.

- 1 × 1 convolutional layer: comparable to a tiny filter, this layer examines a tiny percentage of the input information. As the name '1 × 1' indicates, it uses small filters with 1 × 1 pixel sizes. This layer aids in computing resource conservation by limiting the number of attributes or channels in the data.
- 3 × 3 convolutional layer: to find complex features and patterns in the information, the 3 × 3 convolutional layer uses larger 3 × 3 filters. It operates based on the fewer channels produced by the previous 1 × 1 convolutional layer.

- 1×1 convolutional layer: following the 3 × 3 convolution, the 1x1 convolutional layer does one more round of 1 × 1 convolution. The information's representation is enriched and given an additional purpose by this extra step, which adds additional attributes.

3.3.1 Three-block bottleneck layered model

Then, two bottleneck blocks are included in parallel, including a batch normalisation (BA) layer, ReLU activation layer, convolution layer of depth 64, stride of 1, and filter size of 1 × 1. Subsequently, two bottleneck blocks of a BN layer, a 1-stride filter size, a ReLU activation layer, and a 64-depth convolution layer are added parallel to every bottleneck block. Next, a convolution layer with a depth of 64, a ReLU activation layer, a filter size of 3 × 3, and a stride of 1 is added to this block, after which comes a second BN layer. The ReLU activation and max pooling layers are added as a convolution layer with a depth of 64, a stride of 2, and a filter size of 3 × 3. In the same way, a convolution layer and an additional ReLU activation layer come after introducing a third batch normalising layer. After two more blocks are added, this sequence is repeated.

Then, a ReLU activation layer is added, followed by a filter size of 3 × 3, a convolution layer with a depth of 1,024, and a stride of 2. Subsequently, a ReLU activation layer is added, and another convolution layer with the following parameters is added: filter size = 3 × 3, depth = 2,048, and stride = 2. A softmax layer, a global average pool layer, and a fully connected layer complete the network. The residual system of the three-block bottleneck has 15.9 million learned parameters.

3.3.2 Four-block bottleneck layered model

Three of these bottleneck blocks are stacked to create the neural network that is referred to as having 'four bottleneck blocks'. Every block adheres to the structure mentioned above. The original input dimensions in this network design are 227 × 227 × 3. The first convolutional layer has a filter with a size of 3 × 3, a depth of 32, and a stride of 2, and is followed by an activation layer with ReLU. Next, a maxpooling layer with a stride of 1 and a 3 × 3 filter is applied. Then, a bottleneck block consisting of a ReLU activation layer, a 1-stride filter, a BN layer, a 64-depth convolution layer, and a filter size 1 × 1 is added.

Subsequently, this block received the addition of a second batch normalising layer, a ReLU activation layer, a filter size of 3 × 3, and a convolution layer with a depth of 64. Then, a block consisting of a convolution layer, a ReLU activation layer, and a BN layer is added. ReLU activation and convolution layers are added after adding a second batch normalising layer. In the same way, a convolution layer and an additional ReLU activation layer are integrated with a third BN layer. After two more blocks are added, this sequence is repeated.

In summary, the network is completed by incorporating a softmax layer, a fully connected layer, and a global average pool layer. The four-block bottleneck residual model has a total of 25.1 million learned parameters. The approach becomes scalable and applicable to real-world scenarios due to its efficient processing of massive audio datasets. The model can also extract discriminative features relevant to various emotional expressions since the convolutional layers in DBR-CNNs are excellent at capturing local

temporal dependencies within the audio signals. Optimising the hyper-parameters enhances the classification performance by employing the COA.

3.4 Optimisation

The COA is a proposed algorithm we mathematically model in this section. The COA has several advantages when used for hyper-parameter improvement or optimisation in SER tasks. Because of its innovative methodology, which takes inspiration from cooperative foraging behaviours found in nature, can efficiently explore and exploit the search space, resulting in faster convergence and better performance. Coati, an algorithm designed specifically for optimisation tasks, is helpful for fine-tuning complex models utilised for SER since it efficiently navigates high-dimensional parameter spaces. It may effectively exploit attractive regions while thoroughly examining prospective solutions by balancing exploitation and exploration. This ultimately leads to more reliable and accurate speech-audio emotion identification systems by improving generalisation, decreasing computational costs, and improving model performance.

3.4.1 Inspiration and behaviours of coatis

Members of the Procyonidae family's Nasua and Nasuella genera include coatis, also called coatimundis. Native to Mexico, Central America, South America, and the Southwest of the USA, they are diurnal animals. These dimensions apply to the coatis and white-nosed species found in South America. Smaller than the other, these are called mountain coatis. As omnivores, coatis consume both invertebrates like tarantulas and tiny vertebrate prey like lizards, crocodile eggs, rodents, and bird eggs. A green iguana is one of the coatis's favourite snacks. Coatis chase iguanas in bunches because they are giant reptiles frequently seen in trees.

While some coatis rapidly attack the iguana, others climb trees and intimidate it into jumping to the ground. Nevertheless, coatis are vulnerable to predator attacks. The coati is preyed upon by various animals, including maned wolves, anacondas, dogs, foxes, tayras, jaguarundis, and jaguars.

3.4.2 Initialisation

The values for the decision variables are determined by each coati's location inside the search space. Therefore, coatis' place in the COA indicates a potential fix for the issue. Using equation (5), the coatis's initial position in the search space is randomly determined at the start of the COA implementation.

$$X_i : x_{i,j} = lb_j + r \cdot \left(ub_j - lb_j\right), i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, m \tag{5}$$

X_i represents the i^{th} coati's position in the search space, x_i and j denote the decision variable's value, r is a natural number chosen at random within the interval (0, 1), and lb_j and ub_j, respectively, denote the decision variable's lower and upper bounds.

The following population matrix, or matrix X, is used to depict the coatis population in the COA mathematically.

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{bmatrix}_{N\times m} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,j} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \cdots & x_{i,1} & \cdots & x_{i,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N,1} & \cdots & x_{N,j} & \cdots & x_{N,m} \end{bmatrix}_{N\times m} \tag{6}$$

N is the total number of coatis (population size), m is the number of dimensions (features) that define each individual coati in the population. Equation (6) is used to display these values.

$$X = \begin{bmatrix} F_1 \\ \vdots \\ F_i \\ \vdots \\ F_N \end{bmatrix}_{N\times 1} = \begin{bmatrix} F(X_1) \\ \vdots \\ F(X_i) \\ \vdots \\ F(X_N) \end{bmatrix}_{N\times 1} \tag{7}$$

where $F_i = F(X_i)$ represents the fitness value of the i^{th} coati, calculated by applying the objective function, $F(X)$ to its corresponding position X_i. Every algorithm iteration updates the candidate solutions, which means that every iteration also updates the best member of the population.

3.4.3 Mathematical function of COA

Modelling two of the coatis' natural behaviours forms the basis for upgrading the coati's position in the COA. Among these behaviours are:

3.4.3.1 Exploration phase

The method involves a pack of coatis scaling the tree to get close to an iguana and frighten it. More coatis wait behind a tree when the iguana eventually hits the ground. The coatis hunt down and attack the iguana once it has fallen to the ground. Additionally, it's thought that some coatis climb the tree while others wait for the iguana to drop to the ground. Equation (8) numerically mimics the location of the coatis rising from the tree.

$$X_i^P 1: x_{i,j}^{P1} = x_{i,j} + r \cdot \left(lguana_j - 1 \cdot x_{i,j} \right),\ for\ i = 1, 2, n, \left\lfloor \frac{N}{2} \right\rfloor \text{ and } j = 1, 2, \ldots, m \tag{8}$$

where X_i^{P1} is the position of the ith coati while climbing the tree, $x_{i,j}^{P1}$ is the coordinate of the i^{th} coati in the j^{th} dimension during the climbing phase, $x_{i,j}$ is the coordinate of the i^{th} coati in the j^{th} dimension before updating, r is a random number (typically between 0 and 1) that introduces randomness in movement, $lguana_j$ is the coordinate of the iguana in the j^{th} dimension, N is the total number of coatis in the population, m is the dimensionality of the search space.

The iguana is dropped to the ground and positioned randomly within the search area. equations (9) and (10) simulate the movement of coatis on the ground based on random locations.

$$lguana^G : ?\, lguana_j^G = lb_j + r \cdot \left(ub_j - lb_j\right),\ j = 1, 2, \ldots, m \tag{9}$$

$$X_i^{P1} : x_{i,j}^{P1} = \begin{cases} x_{i,} + r \cdot \left(lguana_j^G - 1 \cdot x_{i,j}\right), & F_{lguana^G} < F_i \\ x_{i,j} + r \cdot \left(x_{i,j} - lguana_j^G\right), & else \end{cases} \tag{10}$$

$$for\ i = \left\lfloor \frac{N}{2} \right\rfloor + 1, \left\lfloor \frac{N}{2} \right\rfloor + 2, \ldots, N\ and\ j = 1, 2, \ldots, m$$

where lb_j is lower boundary of the search space in the j^{th} dimension, ub_j is upper boundary of the search space in the j^{th} dimension, $lguana_j^G$ is the coordinate of the iguana after being dropped to the ground in the j^{th} dimension, F_{lguana^G} is the fitness value of the iguana's new position.

If the new location determined for every coati increases the value of the objective function, then the update mechanism accepts it; if not, the coati stays in its original position. The update condition for the simulated values of i = 1, 2, ..., N is given by equation (11).

$$X_i = \begin{cases} X_i^{P1}, & F_i^{P1} < F_i \\ X_i, & else \end{cases} \tag{11}$$

where F_i is the objective function value at the current position of the i^{th} coati, and F_i^{P1} is the objective function value at the newly proposed position of the i^{th} coati.

3.4.3.2 Exploitation phase

Coatis' natural behaviour when they come across and flee from predators is the basis for a mathematical model. A coati flees its place when a predator attacks it. The fact that coati is in a secure location relative to its current location due to its movements in this strategy indicates that the COA can be exploited in local search.

Equations (12) and (13) create a random position close to each coati's location to replicate this behaviour.

$$lb_j^{local} = \frac{lb_j}{t}, ub_j^{local} = \frac{ub_j}{t}, where\, t = 1, 2, \ldots, T \tag{12}$$

$$X_i^{P2} : x_{i,j}^{P2} = x_{i,j} + (1 - 2r) \cdot \left(lb_j^{local} + r \cdot \left(ub_j^{local} - lb_j^{local}\right)\right), i = 1, 2, \ldots, N, j = 1, 2, \cdots, m \tag{13}$$

$$X_i = \begin{cases} X_i^{P2}, & F_i^{P2} < F_i \\ X_i, & else \end{cases} \tag{14}$$

In equation (12), lb_j^{local} and ub_j^{local} represent the local lower and upper bounds of the search space for the j^{th} dimension. The local bounds are scaled by a factor t, which changes over iterations t = 1, 2, ..., T. This ensures that the search space contracts as the optimisation progresses, refining the solution. In equation (13), X_i^{P2} represents the new candidate position for the i^{th} coati in the j^{th} dimension. The equation generates a random

position around the current location using the local search boundaries. r is a random number between 0 and 1, ensuring stochasticity in movement. (1–2r) introduces a degree of randomness to the movement. The term $lb_j^{local} + r \cdot \left(ub_j^{local} - lb_j^{local}\right)$ defines a local search space around the current position. In equation (14), The coati's position is updated based on the new candidate position X_i^{P2}, F_i^{P2} represents the fitness of the new position, F_i represents the fitness of the current position. If the new position has a better fitness value $\left(F_i^{P2} < F_i\right)$, the coati moves to X_i^{P2}; otherwise, it stays at its current position.

3.4.4 Computation complexity

The first phase's updating of the COA population has a computational cost of O (NmT), where T is the algorithm's iteration count. The complexity of computing the iguana's goal function and random position on the ground in the first phase of COA is equivalent to O(Nm T/2). The computational complexity of the coatis position update method is O(NmT) for the second phase. As a result, O (Nm (1 + 5T/2) is the entire computational complexity of COA.

4 Results and discussion

This section evaluated the SER system's efficacy and compared it with different baseline techniques using a publicly accessible benchmark speech emotions dataset. In this work, we use the four publicly available datasets on speech emotions – EMO-DB, URDU, EMOVO, and SAVEE. The subsequent sub-sections provide a complete overview of the datasets.

4.1 Hyper-parameter configuration

This part presents the system's development findings.

The Table 2 shows the hyperparameter settings.

Table 2 Hyperparameter settings for model training

Hyperparameter	*Value*	*Description*
Batch size	32	Number of training samples processed in one forward and backward pass
Epochs	100	Number of complete passes through the training dataset
Activation function	ReLU, Sigmoid	ReLU prevents vanishing gradients, while Sigmoid is used for binary classification
Dropout rate	0.5-0.1	Regularisation technique to prevent overfitting by randomly dropping neurons
Loss function	Binary cross-entropy	Suitable for binary classification tasks
Learning rate	0.001	Determines the step size in weight updates during training
Optimiser	Adam	Adaptive learning rate optimisation algorithm

4.2 Experimental setup

The system has been developed in a variety of situations. The proposed system boasts a robust environment setup, featuring a Core i5 Gen6 CPU, 8 GB of RAM, and a 4GB GPU, all supported by the versatile Python software.

4.3 Dataset description

The proposed model is evaluated on four publicly available datasets: Emo-DB, URDU, EMOVO, and SAVEE.

4.3.1 Dataset 1 (EMO-DB Berlin emotion dataset)

There are 535 recorded utterances by ten actors (5 male and five female) available in the Berlin emotion database Emo-DB (22). With varying emotions – such as anger, boredom, fear, contempt, happiness, neutrality, and sadness – each actor performed the pre-selected phrases. At a sample rate of 16 kHz, utterances lasting around two to three seconds are included in the Emo-DB.

4.3.2 Dataset 2 (Urdu database)

An audio collection of 400 recordings from 38 speakers (27 men and 11 women) from Urdu TV chat shows can be found in the Urdu database (23). Four basic emotions are the subjects of the data collection: neutral, happy, sad, and angry. Natural and spontaneous emotional clips from conversations between various TV talk show guests are included in this corpus. For research purposes, the dataset is openly accessible.

4.3.3 Dataset 3 (EMOVO (EMOVO corpus: an Italian emotional speech database)

An Italian spoken emotion database called EMOVO (24) recordings six actors—three men and three women – impersonating the seven emotional states of contempt, fear, anger, joy, surprise, sadness, and neutrality. Fourteen sentences represent each emotion, and there are 588 annotated audio recordings in all. The first emotional database for the Italian language, these audio recordings were made in a studio with professional actors and are accessible online.

4.3.4 Dataset 4 (surrey audio-visual expressed emotion database)

To create an automatic emotion identification system, data from the SAVEE database (25) must be gathered. Four recorded male actors exhibiting seven different emotions – annoyance, sadness, neutral, happiness, anger, and disgust – are included in the collection. Overall, 480 British English utterances were available in the sample. Phonetically balanced texts representing each emotion were chosen from the regular TIMIT corpus. High-end audio-visual technology was used in a visual media lab to capture, analyse, and categorise the data. Table 3 shows the details of four emotional datasets.

Table 3 Details of four emotional corpora

Corpus	*Language*	*Utterances*	*Category*	*Positive valence*	*Negative valence*
EMO-DB	German	535	Acted	Boredom, happy, neutral	Anger, fear, disgust, sadness
URDU	Urdu	400	Acted	Neutral, happiness	Anger, sadness
EMOVO	Italian	588	Natural	Neutral, happiness, surprise	Anger, sadness, fear, disgust
SAVEE	English	480	Acted	Neutral, happiness, surprise	Anger, sadness, fear, disgust

4.4 Evaluation metrics

Considered the following performance metrics: precision (P), accuracy (A), recall (R), and F1-score (F) of the proposed approach. According to these metrics:

4.4.1 Accuracy

The accuracy metric is calculated to determine whether SER is accurate.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

4.4.2 Precision

The precise expected positive occurrences to all recognised positive observations ratio is known as precision. The following abilities are characterised by precision:

$$Precision = \frac{TP}{TP + FP} \tag{16}$$

4.4.3 Recall

The recall score shows the recogniser's capacity to find every positive sample. That is the product of FN and TP divided by TP. The following phrases can be used to characterise it:

$$Recall = \frac{TP}{TP + FN} \tag{17}$$

4.4.4 F1-score

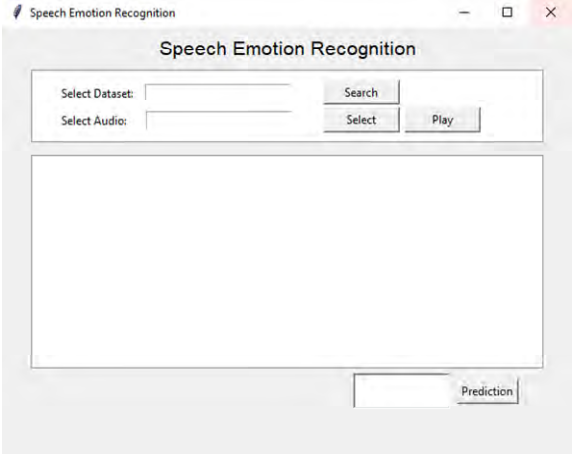
The F1-score determines the harmonic mean of precision and recall.

$$F1\ Score = 2 \times \frac{precision \times recall}{precision + recall} \tag{18}$$

In the context of SER, the confusion matrix helps evaluate how well the model classifies emotions. The four cases assumed in this study are: true positive (TP) → The model correctly predicts an emotion when it is truly present. True negative (TN) → The model

correctly predicts the absence of an emotion. False positive (FP) → the model incorrectly predicts an emotion when it is not present. False negative (FN) → The model fails to detect an emotion that is actually present.

Figure 3 Front page for speech emotion recognition (see online version for colours)

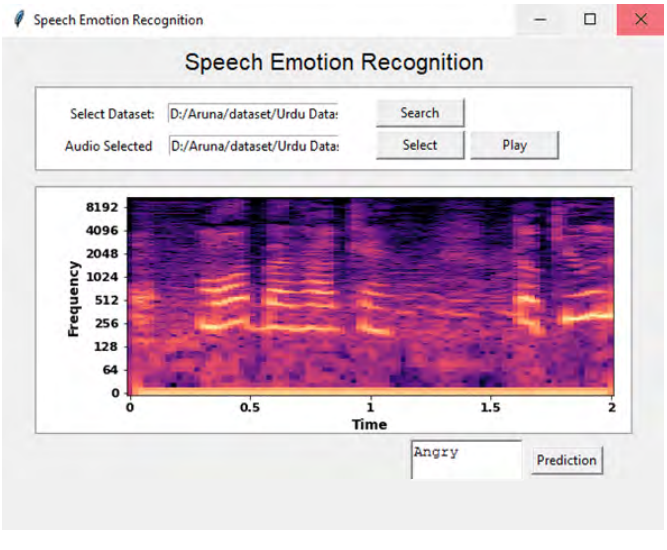


4.5 Implementation outcomes of proposed model

In this part, we demonstrate the outcome of the proposed system as GUI using Python Tkinter.

Figure 3 shows the front page for select datasets and audio files from the device to check the emotions. Figure 4 shows the emotion recognition outcome, the emotion name, and the graph of emotion classification in the given audio signal.

Figure 4 Speech emotion recognition outcome for given input audio (see online version for colours)



4.6 Performance on dataset 1

Figure 5 presents modulation spectral features for various emotions within the EMO-DB dataset. These characteristics are illustrated by modulation spectrograms, which visually represent the patterns of temporal variation of spectral components inside a spectrogram. By examining the modulation spectral patterns associated with various emotions in the EMO-DB dataset, researchers can learn more about how these emotional states appear in the temporal fluctuations of speech spectral components. The development of emotion detection systems based on speech analysis can benefit from this analysis's ability to clarify the unique modulation patterns that distinguish various emotions.

Figure 5 Modulation spectral features for different emotions in the EMO-DB dataset (see online version for colours)

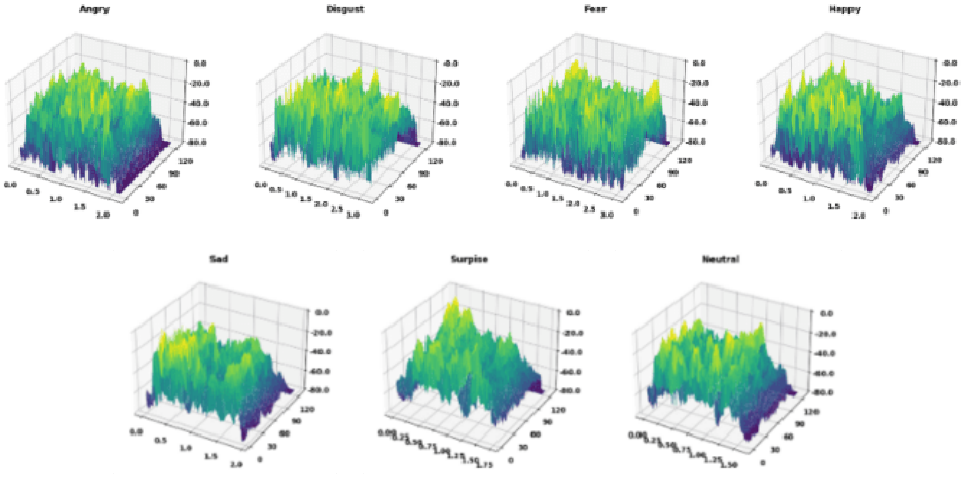


Table 4 Evaluation of the proposed approach for the EMO-DB dataset

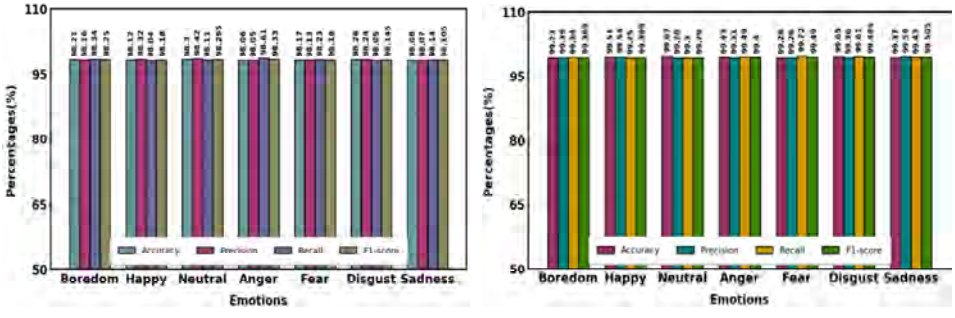
Emotion	*Without optimisation (%)*				*With optimisation (%)*			
	Accuracy	*Precision*	*Recall*	*F1-score*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Boredom	98.21	98.16	98.34	98.25	99.23	99.39	99.34	99.365
Happy	98.12	98.32	98.04	98.18	99.51	99.54	99.25	99.395
Neutral	98.3	98.42	98.11	98.265	99.67	99.28	99.3	99.29
Anger	98.06	98.05	98.61	98.33	99.43	99.31	99.49	99.4
Fear	98.17	98.13	98.23	98.18	99.28	99.26	99.72	99.49
Disgust	98.26	98.24	98.05	98.145	99.65	99.36	99.61	99.485
Sadness	98.08	98.07	98.14	98.105	99.37	99.58	99.43	99.505

Table 4 provides information on the performance of the proposed approach for the EMO-DB dataset. The model performs well without optimisation, with accuracy between 98.06% and 98.3% for various emotions. Comparably high results demonstrate accurate emotion classification for precision, recall, and F1-score. After optimisation, every

emotion category shows a discernible improvement in performance across all parameters. For the Neutral feeling, accuracy increases significantly and reaches as high as 99.67%.

The model's ability to accurately classify emotions is further reinforced by precision, recall, and f1-score improvements. Without optimisation, the model achieves an accuracy of 98.21% and a precision of 98.16%, for example, in the case of the boredom emotion. However, after optimisation, these values dramatically increase to 99.23% and 99.39%, respectively. The model's performance on the EMO-DB dataset is further enhanced via optimisation, as seen by the same increases for other emotions.

Figure 6 The multi-classification outcome of the proposed model is (a) without optimisation and (b) with optimisation on the EMO-DB dataset (see online version for colours)



(a) (b)

Table 5 Comparison of proposed and previous techniques on EMO-DB dataset

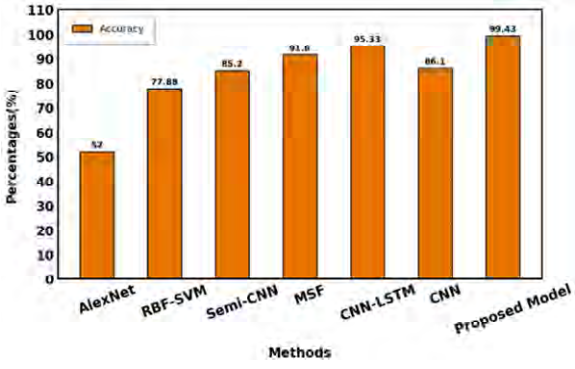
Method	*Accuracy (%)*
AlexNet	52
RBF-SVM	77.88
Semi-CNN	85.2
MSF	91.6
CNN-LSTM	95.33
CNN	86.1
Proposed method	99.43

Figure 6 shows the multi-categorisation outcome of the proposed model with optimisation and without optimisation. Based on accuracy percentages, Table 5 analyses the effectiveness of several methods on the EMO-DB dataset. Achieving 52% accuracy, AlexNet falls short of RBF-SVM's remarkable improvement to 77.88%. Accuracy is further improved by semi-CNN to 85.2% and MSF to 91.6%. CNN-LSTM scores a noteworthy 95.33% accuracy. The CNN model achieves 86.1% accuracy among them.

However, the proposed method stands out with a remarkable accuracy of 99.43%.

This comparison highlights the effectiveness of the proposed method in outperforming earlier approaches in correctly identifying emotions in the EMO-DB dataset, demonstrating its potential for higher performance in emotion detection tasks. Figure 7 illustrates the comparison of proposed and existing models on the EMO-DB dataset.

Figure 7 Comparison of proposed and previous methods on EMO-DB dataset (see online version for colours)



4.7 Performance on dataset 2

Figure 8 shows modulation spectral characteristics in the URDU dataset that correlate to different moods. Speech signal spectral component temporal patterns can be understood using modulation spectrograms. By examining modulation spectral features in relation to distinct emotions found in the URDU dataset, scientists can identify unique patterns of spectral modulation linked to different emotional states.

Figure 8 Modulation spectral features for different emotions in the Urdu dataset (see online version for colours)

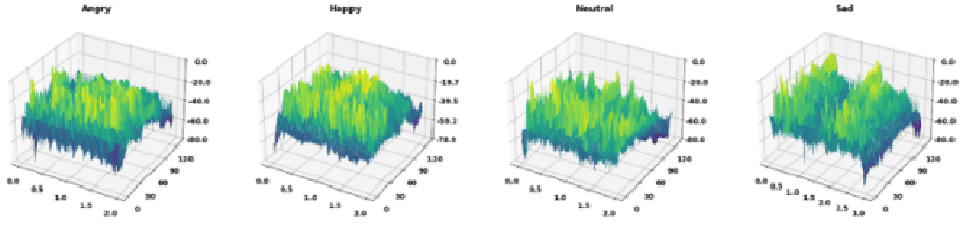


Table 6 presents the results of a proposed framework that focuses on different emotions using the URDU dataset. The effectiveness of the model is assessed both with and without optimisation. The model demonstrates high levels of recall, accuracy, precision, and F1-Score across a range of emotions when it is not optimised. As an illustration, the accuracy for happiness is 97.89%, while the precision, recall, and F1-score are, respectively, 98.34%, 97.94%, and 98.14%. On the other hand, the model shows reasonable performance measures spanning from accuracy to F1-score for emotions such as neutrality, sadness, and anger. The model's performance metrics do, however, significantly improve upon optimisation.

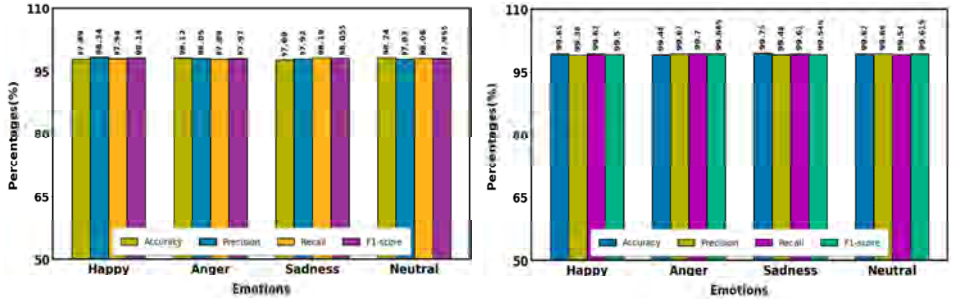
Accuracy, precision, recall, and F1-score all show measurable improvements across all emotions. In proposed model, the accuracy increases considerably to 99.65% in the case of happiness, and the F1-score, precision, and recall reach 99.38%, 99.62%, and 99.5%, respectively. The effectiveness of optimisation in improving the model's performance is demonstrated by the trend's consistency across all other emotions. The proposed approach demonstrates remarkable capability in recognising and categorising

emotions within the URDU dataset. In addition, the model's performance is much improved by optimisation, yielding even more accurate and consistent findings for all emotions assessed.

Table 6 Evaluation of the proposed approach for the URDU dataset

Emotion	*Without optimisation (%)*				*With optimisation (%)*			
	Accuracy	*Precision*	*Recall*	*F1-Score*	*Accuracy*	*Precision*	*Recall*	*F1-Score*
Happy	97.89	98.34	97.94	98.14	99.65	99.38	99.62	99.5
Anger	98.12	98.05	97.89	97.97	99.48	99.67	99.7	99.685
Sadness	97.69	97.92	98.19	98.055	99.75	99.48	99.61	99.545
Neutral	98.24	97.83	98.08	97.955	99.62	99.69	99.54	99.615

Figure 9 The multi-classification outcome of the proposed model is (a) without optimisation and (b) with optimisation on the Urdu dataset (see online version for colours)



(a) (b)

Figure 9 shows the multi-categorisation outcome of the proposed model with optimisation and without optimisation. Table 7 presents a comparative analysis of several methods using an Urdu dataset, showing the differences in their accuracy. Using a mixture of SVM, LR, RF, and NB, the classical approaches achieved accuracies between 75% and 83%. With an accuracy of 82.5%, K-Nearest Neighbours (K-NN) performed slightly better.

Table 7 Comparison of proposed and previous techniques on the URDU dataset

Method	*Accuracy (%)*
NB	76
J48 and NB	75
SVM, LR and RF	83
K-NN	82.5
Proposed method	99.625

The proposed method, however, significantly outperformed all earlier methods and achieved an incredible accuracy of 99.625%.

Figure 10 Comparison of proposed and previous methods on the Urdu dataset (see online version for colours)

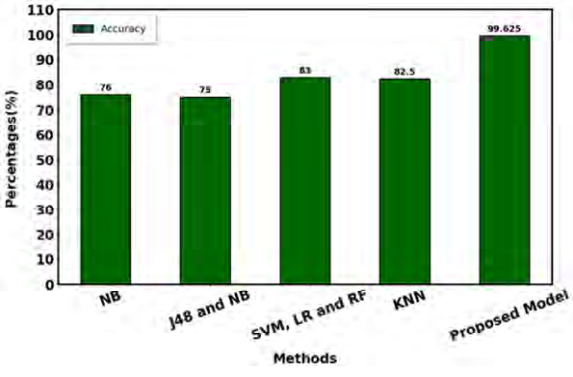
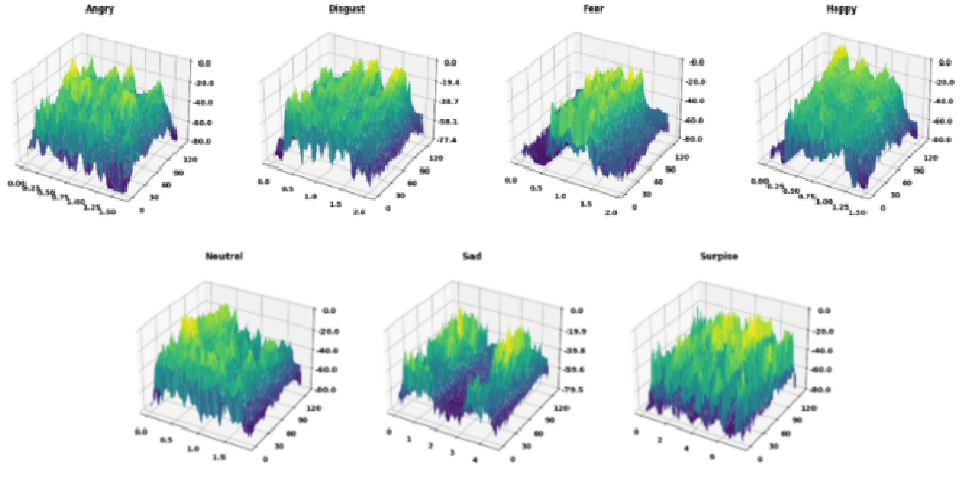


Figure 10 illustrates the comparison of proposed and existing models on the URDU dataset.

Figure 11 Modulation spectral features for different emotions in the EMOVO dataset (see online version for colours)



4.8 Performance on dataset 3

The modulation spectral features of the EMOVO dataset for different emotions are shown in Figure 11. The temporal variations in the spectral components of speech signals are captured by extracting these features. The modulation spectrogram in the picture shows these temporal changes in the spectrum structure. The modulation envelope's low frequencies, in particular, show the slower shifts in the spectral features as a whole, which are thought to represent crucial phonetic information. Through insights into the fundamental mechanics of emotional expression in spoken language, this picture helps explain how various emotions manifest in the modulation features of speech.

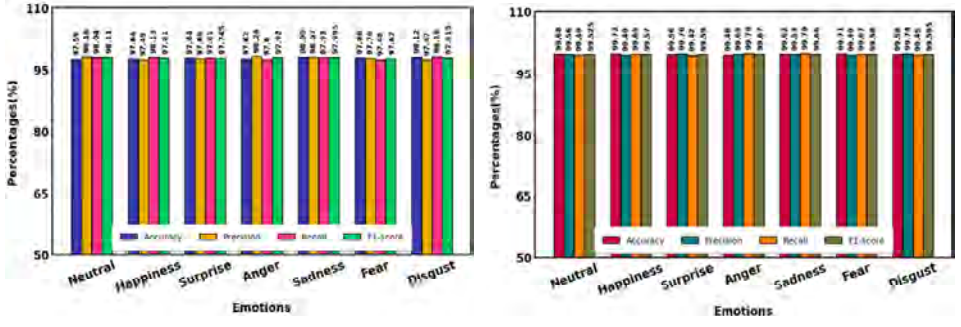
Table 8 shows how the proposed framework performed for the EMOVO dataset and how optimisation affected different emotional categories. The model performs well

across emotions without optimisation, with accuracy ranging from 97.59% to 98.12%. Furthermore, the model's recall, F1-score, and precision measures all show excellent results, demonstrating its ability to identify emotions accurately. After optimisation, performance metrics in every emotional category show a discernible improvement. There is a significant increase in accuracy, with scores as high as 99.72% for fear and 99.71% for happiness.

Table 8 Evaluation of the proposed approach for the EMOVO dataset

Emotion	*Without optimisation*				*With optimisation*			
	Accuracy (%)	*Precision (%)*	*Recall (%)*	*F1-score (%)*	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-score (%)*
Neutral	97.59	98.18	98.04	98.11	99.68	99.56	99.49	99.525
Happiness	97.64	97.49	98.13	97.81	99.72	99.49	99.65	99.57
Surprise	97.84	97.68	97.81	97.745	99.56	99.76	99.42	99.59
Anger	97.62	98.24	97.6	97.92	99.48	99.65	99.78	99.67
Sadness	98.08	98.07	97.92	97.995	99.62	99.53	99.79	99.66
Fear	97.86	97.76	97.48	97.62	99.71	99.49	99.67	99.58
Disgust	98.12	97.47	98.16	97.815	99.58	99.74	99.45	99.595

Figure 12 The multi-classification outcome of the proposed model is (a) without optimisation and (b) with optimisation on the EMOVO dataset (see online version for colours)



(a) (b)

Improvements in precision, recall, and F1-score measures further confirm that optimisation strategies effectively improve the model's predictive power. Overall, the findings demonstrate how optimisation can improve the model's accuracy and ability to identify complex emotional states in the EMOVO dataset.

Figure 12 shows the multi-categorisation outcome of the proposed model with optimisation and without optimisation. Table 9 presents an analysis of the efficacy of several approaches on the EMOVO dataset, with a particular emphasis on accuracy percentages. The accuracy of conventional techniques such as MLP and SVM is 58.58% and 60.40%, respectively. More sophisticated models with accuracies of 53.24% and 69.65%, respectively, are CNN-LSTM and DCNN.

With an incredible accuracy of 99.62%, the proposed approach, however, stands out considerably. This implies that the proposed method performs far better on the EMOVO

dataset than earlier approaches, proving its effectiveness and potential to improve sentiment analysis work.

Table 9 Comparison of proposed and previous techniques on the EMOVO dataset

Method	*Accuracy (%)*
SVM	60.40
MLP	58.58
CNN-LSTM	53.24
DCNN	69.65
proposed method	99.62

Figure 13 Comparison of proposed and previous methods on EMOVO dataset (see online version for colours)

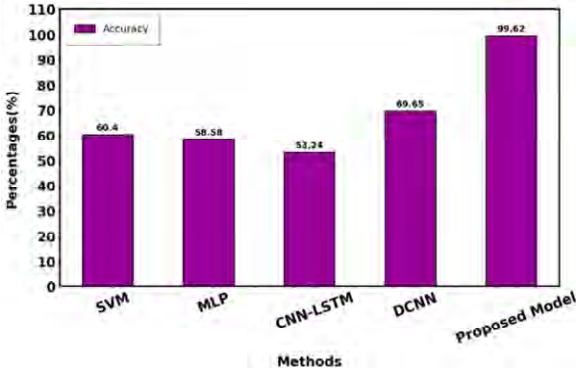


Figure 13 compares proposed and existing models on the EMOVO dataset.

4.9 Performance on dataset 4

The modulation spectral features in the SAVEE dataset that correlate to different emotions are shown in Figure 14. Because of this, the modulation spectrogram is a helpful tool for analysing these fluctuations in spectral properties and identifying emotional states in voice data.

Table 10 analyses the evaluation of the proposed approach with and without optimisation on the SAVEE dataset. With anger scoring 97.84%, disgust scoring 97.69%, fear scoring 98.18%, happiness scoring 97.67%, Neutral scoring 98.06%, sadness scoring 97.94%, and surprise scoring 97.43%, the model exhibits great accuracy while not optimised. Each emotion also shows strong values for F1-score, recall, and precision.

However, with optimisation, performance measurements show a significant improvement, with notable increases in recall, accuracy, precision, and F1-score. Anger, for example, has an accuracy improvement of 99.75%, and comparable improvements are seen for disgust, fear, happy, neutral, sadness, and surprise. These findings demonstrate how optimisation techniques can improve the model's predicted performance and produce more reliable emotion recognition across the SAVEE dataset.

Table 10 Evaluation of the proposed approach for the SAVEE dataset

Emotion	*Without optimisation*				*With optimisation*			
	Accuracy (%)	*Precision (%)*	*Recall (%)*	*F1-score (%)*	*Accuracy (%)*	*Precision (%)*	*Recall (%)*	*F1-score (%)*
Anger	97.84	97.72	97.46	97.59	99.75	99.45	99.65	99.55
Disgust	97.69	98.04	98.12	98.08	99.62	99.62	99.51	99.565
Fear	98.18	97.81	97.67	97.74	99.58	99.51	99.73	99.62
Happy	97.67	97.34	97.55	97.445	99.73	99.55	99.67	99.61
Neutral	98.06	98.25	98.12	98.185	99.59	99.63	99.71	99.67
Sadness	97.94	98.46	97.89	98.175	99.61	99.47	99.59	99.53
Surprise	97.43	97.48	97.73	97.605	99.79	99.57	99.62	99.595

Figure 14 Modulation spectral features for different emotions in the SAVEE dataset (see online version for colours)

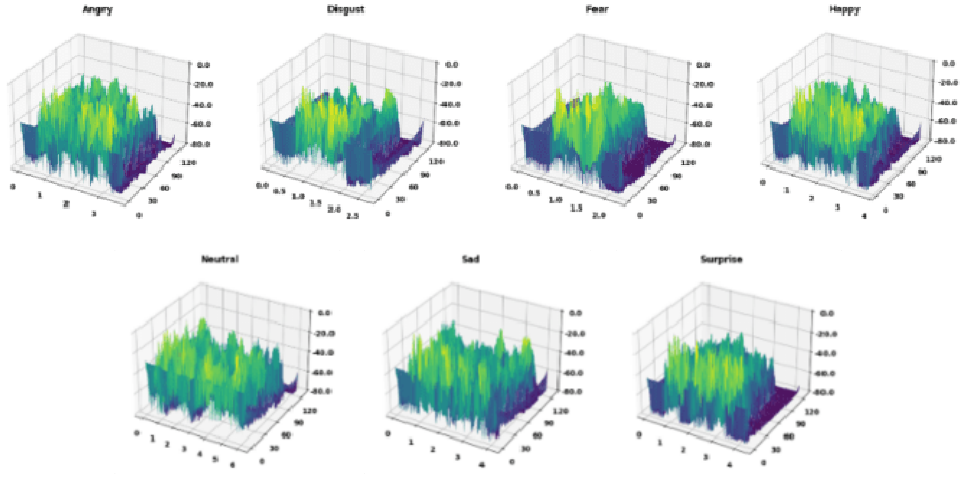
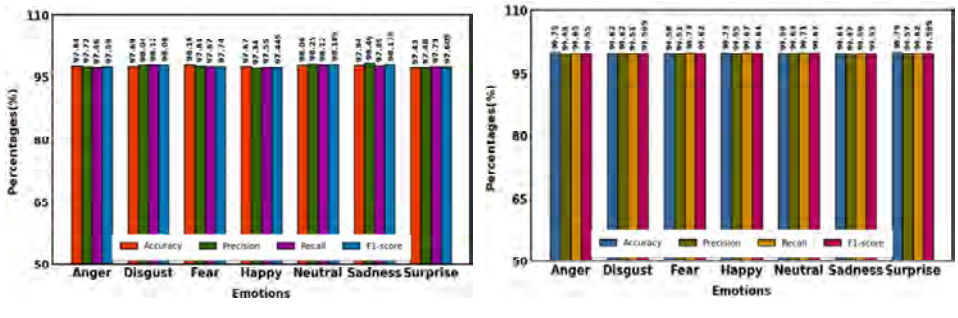


Figure 15 The multi-classification outcome of the proposed model is (a) without optimisation and (b) with optimisation on the SAVEE dataset (see online version for colours)



(a) (b)

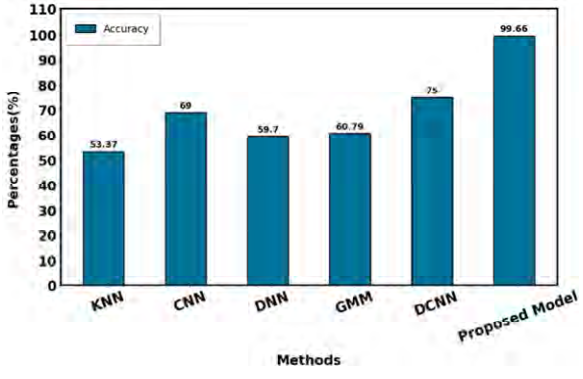
Figure 15 shows the multi-categorisation outcome of the proposed model with optimisation and without optimisation. As expressed in accuracy percentages, Table 11 compares the effectiveness of several methods used on the SAVEE dataset. The accuracy of conventional techniques like Gaussian mixture models (GMM), DNN, and k-nearest neighbours (K-NN) ranged from 53.37% to 60.79%. Advanced techniques like CNN and DCNN improved accuracy by 69% and 75%.

Table 11 Comparison of proposed and previous techniques on the SAVEE dataset

Method	*Accuracy (%)*
K-NN	53.37
CNN	69
DNN	59.7
GMM	60.79
DCNN	75
Proposed method	99.66

Nevertheless, a recently proposed approach with a fantastic accuracy of 99.66% greatly exceeded all previous approaches.

Figure 16 Comparison of proposed and previous methods on the SAVEE dataset (see online version for colours)



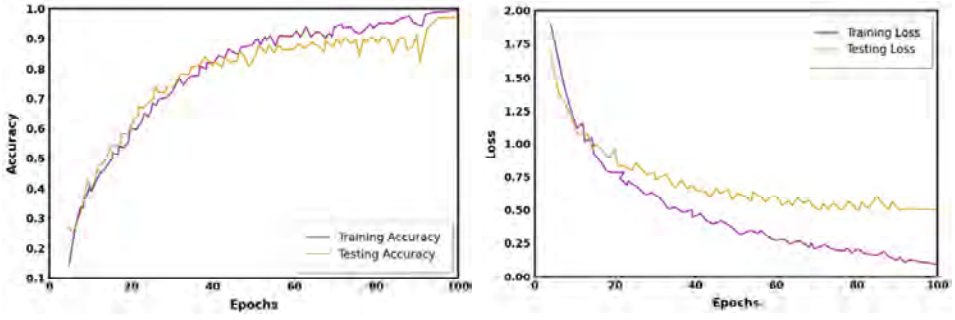
This indicates the proposed method effectively identifies data from the SAVEE dataset and represents a significant achievement in the field. Figure 16 illustrates the comparison of proposed and existing models on the SAVEE dataset.

4.10 Evaluation of training and testing

Validation data is then used to evaluate the trained model's performance to alter hyper-parameters and avoid overfitting.

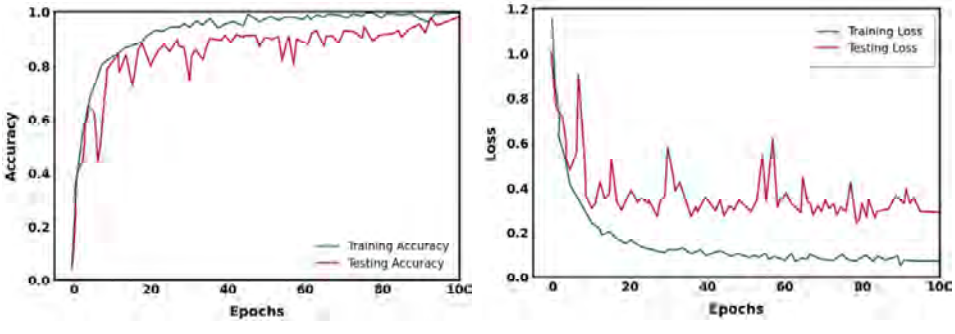
Figures 17, 18, 19, and 20 display a loss value and categorisation accuracy graph as the number of iteration steps increased. The graph shows how the approach covered in this study benefits convergence.

Figure 17 Evaluation performance of training and testing on EMO-DB dataset (a) accuracy (b) loss (see online version for colours)



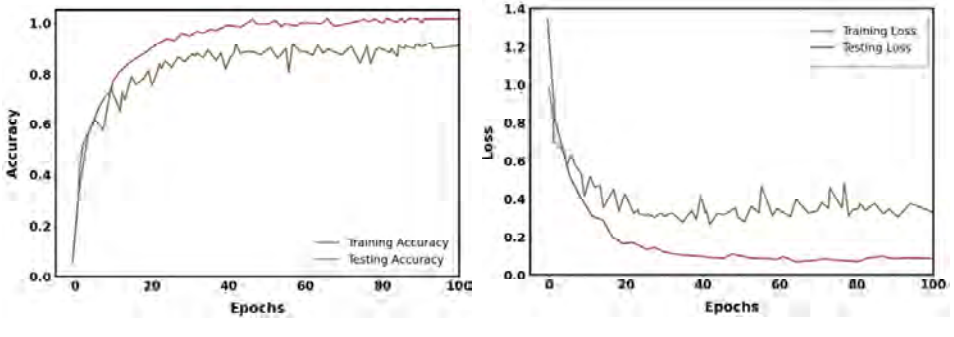
(a) (b)

Figure 18 Evaluation performance of training and testing on Urdu dataset (a) accuracy (b) loss (see online version for colours)



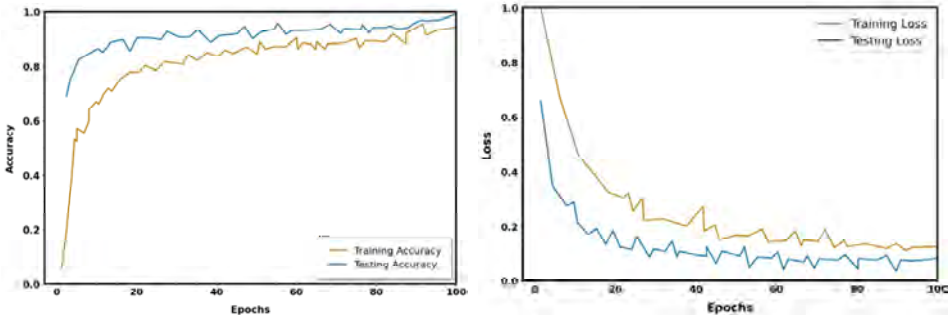
(a) (b)

Figure 19 Evaluation performance of training and testing on EMOVO dataset (a) accuracy (b) loss (see online version for colours)



(a) (b)

Figure 20 Evaluation of training and testing performance on SAVEE dataset (a) accuracy (b) loss (see online version for colours)



(a) (b)

4.11 Overall comparison of proposed method and existing methods in literature

Table 12 provides a comparative analysis of the proposed optimised DBR-CNN against several previously established methods for SER. The comparison focuses on the performance of each approach based on accuracy percentages obtained from various benchmark datasets, including TESS, EMO-DB, RAVDESS, SAVEE, CREMA-D, IEMOCAP, URDU, and EMOVO. The CNN-LSTM-GRU hybrid model, introduced by Ahmed et al. (16), demonstrates strong performance across multiple datasets, achieving high accuracy on TESS (99.46%), EMO-DB (95.42%), RAVDESS (95.62%), SAVEE (93.22%), and CREMA-D (90.47%). This indicates the effectiveness of combining convolutional layers with LSTM and GRU units in capturing both spatial and temporal dependencies in speech signals. However, while CNN-LSTM-GRU performs well, there is still room for improvement, particularly on datasets like CREMA-D and SAVEE, where the accuracy is below 95%.

Xu et al. (17) proposed HD-MFM, which yields an accuracy of 91.25% on EMO-DB and a considerably lower accuracy of 72.02% on the more challenging IEMOCAP dataset. Similarly, the CoSTGA model by Kakuba et al. (18) achieves an accuracy of 75.82% on IEMOCAP, highlighting the difficulty of generalising well on datasets that contain more complex and spontaneous emotional expressions. The LSTM-transformer architecture, introduced by Andayani et al. (19), obtains 85.55% on EMO-DB and 75.62% on RAVDESS, suggesting that while transformers can enhance temporal modelling, the overall performance still lags behind the more modern deep learning methods.

Jothimani and Premalatha (20) leveraged a CNN-based model that achieved 99.6% accuracy on TESS, showing its suitability for that dataset. However, the CNN model underperformed on SAVEE (84.9%) and CREMA (89.9%) compared to newer methods, indicating limitations in its capacity to capture complex emotional patterns in speech. In contrast, the proposed optimised DBR-CNN approach consistently outperforms the aforementioned models across multiple datasets. It achieves 99.43% on EMO-DB, 99.625% on URDU, 99.62% on EMOVO, and 99.66% on SAVEE. These results highlight the robustness and superior generalisation capability of the proposed model across datasets with varying linguistic and acoustic characteristics.

Table 12 Overall comparison of proposed and previous methods in the literature

Reference	*Method*	*Dataset*	*Accuracy (%)*
Ahmed et al. (16)	CNN-LSTM-GRU	TESS	99.46
		EMO-DB	95.42
		RAVDESS	95.62
		SAVEE	93.22
		CREMA-D	90.47
Xu et al. (17)	HD-MFM	EMO-DB	91.25
		IEMOCAP	72.02
Kakuba et al. (18)	CoSTGA	IEMOCAP	75.82
Andayani et al. (19)	LSTM-transformer	RAVDESS	75.62
		Emo-DB	85.55
Jothimani and Premalatha (20)	CNN	RAVDESS	92.6
		CREMA	89.9
		SAVEE	84.9
		TESS	99.6
Proposed method	Optimised DBR-CNN	EMO-DB	99.43
		URDU	99.625
		EMOVO	99.62
		SAVEE	99.66

The proposed optimised DBR-CNN model offers several advantages over previous methods in SER. It achieves superior accuracy across multiple benchmark datasets, consistently exceeding 99%, demonstrating its exceptional generalisation capability across varied linguistic and emotional speech data. By leveraging the SEResNeXt-101 technique for feature extraction, the model efficiently captures discriminative temporal and spectral features while reducing computational complexity compared to traditional CNN or hybrid models like CNN-LSTM-GRU. The incorporation of bottleneck residual blocks in the DBR-CNN architecture enhances gradient flow, enabling deeper model training and improved feature learning without degradation. Additionally, the integration of the COA optimises hyperparameters and further refines the feature space, leading to enhanced recognition and classification performance. The model also benefits from robust preprocessing techniques, such as noise reduction and normalisation, which improve its resilience to noisy or real-world speech inputs, addressing a common limitation found in previous methods. Overall, the proposed approach outperforms existing models by delivering higher accuracy, better feature representation, scalability, and adaptability, making it a highly effective solution for real-world SER applications.

The proposed model significantly outperforms SOTA methods in SER across multiple datasets. Table 13 presents a comparative analysis of various SER approaches based on different feature extraction techniques and datasets. Traditional methods such as EmoBox and FLUDA, which rely on speech representations, exhibit relatively low accuracy, with EmoBox achieving 48.12% on RAVDESS and 49.30% on SAVEE, while FLUDA reaches only 56.8% on EMODB. Similarly, methods integrating speech and text

representations, such as MDAT, show moderate performance, achieving 85.51% on EMOVO but only 42.48% on EMODB.

Table 13 Comparison with the state-of-the-arts methods

Method	*Features*	*Dataset*	*Accuracy (%)*
EmoBox	Speech representations	RAVDEES	48.12
		SAVEE	49.30
		MELD	51.42
MDAT	Speech representation and text embedding	EMOVO	85.51
		EMODB	42.48
FLUDA	Speech representation	EMODB	56.8
Ensemble	MFCC, low-level acoustic features	URDU	62.5
VACNN + BOVE	Log-mel spectrogram	EMODB	86.92
		SAVEE	75
		RAVDESS	83.33
ADRNN	Log-mel spectrogram	EMODB	63.84
GAN-SVM	Low-level acoustic features	EMODB	63.25
		SAVEE	55.12
		EMOVO	61.8
Proposed model	SEResNeXt-101	URDU	99.625
		EMOVO	99.62
		SAVEE	99.66

More advanced models like VACNN + BOVE and ADRNN utilise log-mel spectrograms for feature extraction, yielding higher accuracy rates, particularly VACNN + BOVE, which achieves 86.92% on EMODB and 83.33% on RAVDESS. However, approaches like GAN-SVM and Ensemble, which rely on low-level acoustic features and MFCCs, still struggle with lower accuracy, ranging from 55.12% to 62.5%. In contrast, the proposed model, which leverages SEResNeXt-101 for feature extraction, achieves remarkable accuracy across all datasets, with 99.625% on URDU, 99.62% on EMOVO, and 99.66% on SAVEE. This demonstrates the effectiveness of the deep learning-based approach in capturing complex speech patterns and emotional variations more efficiently than conventional and hybrid methods. The results highlight the robustness and superior generalisation capabilities of the proposed model, making it a promising solution for real-world SER applications

4.12 Computational time complexity

Table 14 compares the running times of different approaches, including the proposed and earlier approaches. While Xu et al. used HD-MFM and took 0.548 ms, Ahmed et al. used CNN-LSTM-GRU, which had a running duration of 0.265 ms. Kakuba et al. used CoSTGA and achieved a running time of 0.452 ms.

Table 14 Comparison of proposed and previous method's running time

Reference	*Method*	*Running time (ms)*
Ahmed et al. (16)	CNN-LSTM-GRU	0.265
Xu et al. (17)	HD-MFM	0.548
Kakuba et al. (18)	CoSTGA	0.452
Andayani et al. (19)	LSTM-transformer	0.374
Jothimani and Premalatha (20)	CNN	0.592
Proposed method	Optimised DBR-CNN	0.196

Jothimani and Premalatha used CNN with a running time of 0.592 ms, while Andayani et al. selected LSTM-Transformer with a running time of 0.374 ms. Optimised DBR-CNN, the proposed method, achieves a much shorter running time of 0.196 ms and exceeds all previous methods in terms of efficiency. This implies that, compared to current methods, the proposed approach offers considerable gains in processing speed, indicating that it is a promising strategy for upcoming uses.

4.13 Limitations of the proposed model and future scopes

While the proposed DBR-CNN combined with the COA demonstrates outstanding performance in SER, achieving SOTA accuracy across multiple datasets, several limitations still remain. One notable limitation is the dependency on a limited number of benchmark datasets (URDU, EMO-DB, EMOVO, and SAVEE), which may not fully capture the diversity and variability of real-world speech emotions across different languages, accents, and speaking styles. Additionally, while SEResNeXt-101 effectively reduces complexity and extracts meaningful features, the model's computational demand during training and inference stages can still be high, which could hinder its deployment in resource-constrained or real-time applications. Another challenge is the limited generalisation ability when dealing with noisy or highly spontaneous speech from unconstrained environments, where performance may degrade without further fine-tuning.

For future work, we plan to address these limitations by expanding the system to include a wider range of multilingual and real-world datasets, including in-the-wild and cross-corpus datasets, to enhance robustness and generalisation. Moreover, we aim to explore lightweight model variants or knowledge distillation techniques to reduce computational overhead while retaining high accuracy, making the model more suitable for edge devices and real-time applications. Another avenue for future research includes incorporating multimodal data (e.g., facial expressions, physiological signals) alongside speech to develop a more comprehensive emotion recognition framework. Finally, investigating adaptive noise filtering and domain adaptation techniques could further improve model resilience in diverse and noisy environments.

5 Conclusions

This study introduced a novel deep learning-based approach for SER, addressing the complexity of feature extraction and accurate emotion classification. The proposed model

integrated the DBR-CNN framework with the SEResNeXt-101 technique for feature extraction, along with the COA to enhance recognition performance. A comprehensive evaluation was conducted using four widely used datasets – URDU, EMO-DB, EMOVO, and SAVEE – demonstrating the superiority of the proposed method over existing approaches. The results indicate that the model achieves a significantly higher accuracy rate, reducing computational time and improving emotion recognition in challenging conditions. Compared to conventional SER techniques, the proposed approach effectively captures both short-term and long-term speech dependencies while minimising redundant computations. The ensemble structure ensures robustness in recognising complex emotions across different languages and speaker variations. The experimental results confirm that the model surpasses SOTA techniques, achieving an average recognition accuracy of 99.43% for URDU, 99.625% for EMO-DB, 99.62% for EMOVO, and 99.66% for SAVEE. The findings of this research contribute to the advancement of SER by providing an efficient and scalable model that can be applied to various HCI scenarios, including virtual assistants, emotional health monitoring, and intelligent communication systems.

Declarations

Availability of data and material: data will be availed when requested.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdulmohsin, H.A. (2021) 'A new proposed statistical feature extraction method in speech emotion recognition', *Computers and Electrical Engineering*, Vol. 93, No. 1, p.107172.

Ahmed, M.R., Islam, S., Islam, A.M. and Shatabda, S. (2023) 'An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition', *Expert Systems with Applications*, Vol. 218, No. 1, p.119633.

Ancilin, J. and Milton, A. (2021) 'Improved speech emotion recognition with Mel frequency magnitude coefficient', *Applied Acoustics*, Vol. 179, No. 1, p.108046.

Andayani, F., Theng, L.B., Tsun, M.T. and Chua, C. (2022) 'Hybrid LSTM-transformer model for emotion recognition from speech audio files', *IEEE Access*, Vol. 10, No. 1, pp.36018–36027.

Bhangale, K. and Kothandaraman, M. (2023) 'Speech emotion recognition based on multiple acoustic features and deep convolutional neural network', *Electronics*, Vol. 12, No. 4, p.839.

Chen, Z., Li, J., Liu, H., Wang, X., Wang, H. and Zheng, Q. (2023) 'Learning multi-scale features for speech emotion recognition with connection attention mechanism', *Expert Systems with Applications*, Vol. 214, No. 1, p.118943.

Emo-Db Berlin Emotion Dataset [online] https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb?select=wav.

EMOVO Corpus: An Italian Emotional Speech Database [online] https://www.kaggle.com/datasets/sourabhy/emovo-italian-ser-dataset.

Gerczuk, M., Amiriparian, S., Ottl, S. and Schuller, B.W. (2021) 'Emonet: a transfer learning framework for multi-corpus speech emotion recognition', *IEEE Transactions on Affective Computing*, No. 1.

Jothimani, S. and Premalatha, K. (2022) 'MFF-SAug: multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network', *Chaos, Solitons and Fractals*, Vol. 162, No. 1, p.112512.

Kakuba, S., Poulose, A. and Han, D.S. (2022) 'Deep learning-based speech emotion recognition using multi-level fusion of concurrent features', *IEEE Access*, Vol. 10, No. 1, pp.125538–125551.

Krishnan, P.T., Joseph Raj, A.N. and Rajangam, V. (2021) 'Emotion classification from speech signal based on empirical mode decomposition and non-linear features: speech emotion recognition', *Complex and Intelligent Systems*, Vol. 7, No. 1, pp.1919–1934.

Kwon, S. (2021) 'MLT-DNet: speech emotion recognition using 1D dilated CNN based on multi-learning trick approach', *Expert Systems with Applications*, Vol. 167, No. 1, p.114177.

Li, D., Liu, J., Yang, Z., Sun, L. and Wang, Z. (2021) 'Speech emotion recognition using recurrent neural networks with directional self-attention', *Expert Systems with Applications*, Vol. 173, No. 1, p.114683.

Liu, Z.T., Han, M.T., Wu, B.H. and Rehman, A. (2023) 'Speech emotion recognition based on convolutional neural network with attention-based bidirectional long short-term memory network and multi-task learning', *Applied Acoustics*, Vol. 202, No. 1, p.109178.

Shahin, I., Alomari, O.A., Nassif, A.B., Afyouni, I., Hashem, I.A. and Elnagar, A. (2023) 'An efficient feature selection method for Arabic and English speech emotion recognition using Grey wolf optimizer', *Applied Acoustics*, Vol. 205, No. 1, p.109279.

Singh, P., Srivastava, R., Rana, K.P.S. and Kumar, V. (2021) 'A multimodal hierarchical approach to speech emotion recognition from audio and text', *Knowledge-Based Systems*, Vol. 229, No. 1, p.107316.

Sun, C., Li, H. and Ma, L. (2023) 'Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network', *Frontiers in Psychology*, Vol. 13, No. 1, p.1075624.

Surrey Audio-Visual Expressed Emotion Database [online] https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee.

URDU Dataset [online] kaggle.com/datasets/hazrat/urdu-speech-dataset?select=files.

Xu, X., Li, D., Zhou, Y. and Wang, Z. (2022) 'Multi-type features separating fusion learning for speech emotion recognition', *Applied Soft Computing*, Vol. 130, p.109648.

Yildirim, S., Kaya, Y. and Kılıç, F. (2021) 'A modified feature selection method based on metaheuristic algorithms for speech emotion recognition', *Applied Acoustics*, Vol. 173, p.107721.

Zehra, W., Javed, A.R., Jalil, Z., Khan, H.U. and Gadekallu, T.R. (2021) 'Cross corpus multi-lingual speech emotion recognition using ensemble learning', *Complex and Intelligent Systems*, pp.1–10.

Zhang, S., Tao, X., Chuang, Y. and Zhao, X. (2021) 'Learning deep multimodal affective features for spontaneous speech emotion recognition', *Speech Communication*, Vol. 127, pp.73–81.