RESEARCH PAPER

BGWO-Based Multidomain Feature Optimization for Multilingual Speech Emotion Recognition

[1]Amarja Adgaonkar, [2]Dr. Jayshree Jain and [3]Dr. Umesh Kulkarni

[1,2]Department of Computer Engineering, Pacific Academy of Higher Education and Research University, Udaipur, India
[3]Department of Computer Engineering, Vidyalankar Institute of Technology, Mumbai, India
[1]adgaonkar123@gmail.com, [2]drjayshreejain@gmail.com and [3]umesh.kulkarni@vit.edu.in

Received: 24th April, 2026; Revised: 20th May 2026; Accepted: 30th June, 2026; Available Online: 10th July, 2026

ABSTRACT

Speech Emotion Recognition (SER) plays an important role in human–computer interaction, affective computing, healthcare, and intelligent conversational systems. However, multilingual SER remains challenging due to language variability, emotional ambiguity, and the presence of redundant high-dimensional acoustic features. This work proposes a BGWO-based multidomain feature optimization framework for multilingual SER using Bengali, Hindi, and Telugu emotional speech datasets for anger, disgust, fear, happy, neutral, and surprise emotions using Random Forest (RF) and Multilayer Perceptron (MLP) classifiers , ensuring robustness across both ensemble-based and neural-network learning paradigms. Initially, a baseline SER system was developed using Mel Frequency Cepstral Coefficients (MFCC) and Delta features . To improve emotional discriminative capability, a multidomain handcrafted feature representation consisting of time-domain, spectral, cepstral, and time–frequency features is introduced. Further, Binary Grey Wolf Optimization (BGWO) is employed to perform optimal feature subset selection and dimensionality reduction. Experimental analysis demonstrates that multidomain features significantly outperform MFCC-based representations across all languages. The proposed BGWO framework reduces the original 192-dimensional feature space by approximately 31–36% while maintaining competitive SER performance. The best Phase 2 accuracies achieved were 81% for Bengali, 82% for Hindi, and 78% for Telugu. Experimental findings further indicate that optimization behavior varies across languages and dataset scales. The proposed framework demonstrates the effectiveness of multidomain feature engineering and optimization for efficient multilingual speech emotion recognition.

How to cite this article: Adgaonkar A, Jain J, Kulkarni U., BGWO-Based Multidomain Feature Optimization for Multilingual Speech Emotion Recognition. Int J Drug Deliv Technol. 2026;16(76s): 573; DOI: 10.25258/ijddt.16.76s.52
Source of support: Nil.
Conflict of interest: None

1. INTRODUCTION

Speech Emotion Recognition (SER) has emerged as an important research area in affective computing and human–computer interaction due to its ability to automatically identify emotional states from speech signals. SER systems are increasingly being utilized in various real-world applications such as virtual assistants, intelligent tutoring systems, healthcare monitoring, call-center analytics, and conversational artificial intelligence.[1] Human speech contains rich emotional information conveyed through acoustic, spectral, prosodic, and temporal variations, making SER an essential component for developing emotionally intelligent computing systems. [2] Despite significant progress in recent years, achieving robust and generalized emotion recognition across multiple languages remains a challenging problem[3] .

Traditional SER systems primarily rely on Mel Frequency Cepstral Coefficients (MFCCs) because of their effectiveness in representing perceptual speech characteristics. Several studies have demonstrated satisfactory performance using MFCC-based representations combined with machine learning and deep learning classifiers. However, MFCC features mainly capture cepstral information and may not sufficiently represent complex emotional variations present in speech. Emotional speech signals often exhibit nonlinear temporal, spectral, and prosodic characteristics that cannot be fully captured using cepstral features alone. Consequently, SER systems based solely on MFCC features may suffer from limited discriminative capability, particularly in multilingual and emotionally diverse environments.

To overcome these limitations, recent SER research has increasingly explored multidomain handcrafted feature representations that combine complementary acoustic information from multiple feature domains. Such representations typically include time-domain features, spectral descriptors, cepstral coefficients, and time–frequency characteristics. The integration of multidomain features enables improved emotional representation and enhances classifier learning capability. However, multidomain feature extraction often leads to high-dimensional feature spaces containing redundant and less informative features. The presence of redundant features increases computational complexity and may negatively affect classifier generalization performance.

**Author for Correspondence: adgaonkar123@gmail.com*

Feature optimization and feature selection techniques have therefore become important components in modern SER frameworks. Metaheuristic optimization algorithms are widely employed for selecting discriminative feature subsets while reducing dimensionality. Among these approaches, Grey Wolf Optimization (GWO) has attracted attention due to its strong exploration–exploitation balance and simple implementation mechanism. Binary Grey Wolf Optimization (BGWO), an extension of GWO for binary search spaces, provides an effective mechanism for selecting optimal feature subsets in high-dimensional classification problems. By eliminating redundant features, BGWO can improve classification robustness while simultaneously reducing computational overhead. [30].

In addition to feature-related challenges, multilingual SER introduces further complexities due to variations in phonetics, pronunciation patterns, speaking styles, and emotional expression across languages. Most existing SER studies primarily focus on single-language datasets, limiting the generalizability of proposed approaches. Comparative multilingual investigations involving Indian regional languages remain relatively limited in current literature. Therefore, developing efficient multilingual SER frameworks capable of handling diverse emotional speech patterns is an important research direction.

Motivated by these challenges, this work proposes a BGWO-based multidomain feature optimization framework for multilingual speech emotion recognition using Bengali, Hindi, and Telugu emotional speech datasets. Initially, a baseline SER framework was developed using MFCC and Delta features with Multi-Layer Perceptron (MLP) and Random Forest (RF) classifiers. Subsequently, a multidomain handcrafted feature representation consisting of time-domain, spectral, cepstral, and time–frequency features is introduced to improve emotional discriminative capability. Further, Binary Grey Wolf Optimization is employed to perform optimal feature subset selection and dimensionality reduction. Experimental evaluation is conducted across multilingual datasets to analyze the effectiveness of multidomain features and optimization-based feature selection for SER performance enhancement.

The major contributions of this work are summarized as follows:

1. A multilingual SER framework is developed using Bengali, Hindi, and Telugu emotional speech datasets.
2. A comparative analysis between MFCC-based and multidomain handcrafted feature representations is performed for multilingual SER.
3. The performance of MLP and RF classifiers is investigated across multiple feature representations and languages.
4. A BGWO-based feature optimization framework is proposed for selecting discriminative multidomain feature subsets and reducing feature dimensionality.
5. The proposed framework analyzes language-dependent optimization behavior and its impact on multilingual SER performance.

2. RELATED WORK

A multilingual Speech Emotion Recognition (SER) system is essential in countries such as India, which is characterized by a diverse linguistic landscape. The body of literature on automatic SER systems in Indian languages is limited compared to that in foreign languages. A primary challenge in the context of Indian languages is the inadequacy of databases, specifically the availability of established and publicly accessible data resources in the form of acted, simulated, and natural databases. Given the variations in dialects, accents, speaking styles, and speaking rates, as well as the cultural and environmental contexts in which speakers' express emotions.

According to [4] the selection of appropriate features and classifiers is essential for achieving accurate Speech Emotion Recognition (SER). However, the trajectory of emotion recognition remains ambiguous for many Indian languages.

The SER system is a collection of methodologies for analyzing and classifying speech data to discover the embedded emotions [2]. In automatic SER designing, the first requirement of building an SER system is a suitable speech dataset having different emotional states. There are three kinds of databases that are used for the development of speech emotion recognition systems: actor (simulated), elicited (induced), and natural emotional speech databases. Simulated speech corpora are collected from experienced and trained artists. Elicited or induced speech corpora are collected by simulating artificial emotional situations without the speaker's knowledge. Naturally available real-world speech is collected from various sources such as call center data, etc. [4].

Once the data is collected, the raw speech data goes through some pre-processing techniques such as noise reduction, silence removal, framing, windowing, and normalization for enhancing the speech signal. The approach for speech emotion recognition (SER) primarily comprises two phases known as the feature extraction and feature classification phases [6].

In the field of speech processing, researchers have derived several features, such as source-based excitation features [4], prosodic features [7][8][2][9], voice quality features [10], combinations of features [11][12][13][14][15][16][17][18][19], phase-based cepstral coefficients [3].

The most widely used spectral feature in automatic speech recognition is the Mel Frequency Cepstral Coefficient (MFCC) [20]. [21] computed MFCC, delta MFCC, delta-delta MFCC, and total log-energy features from each frame using Mel-frequency filter banks for recognition of Assamese (native language). In machine recognition of emotion, [22],[23] tried MFCC features and intensity and pitch values extracted from each frame of the samples.

[12][24][25][26][27] used MFCC features for Marathi, Tamil, Malayalam, Indian English, Kannada, and Bengali language SER, respectively.

Gammatone Mel Frequency Cepstral Coefficient (GMFCC) deals with basilar membrane displacement obtained from the gammatone filter, and it is useful for recognizing gender from emotional speech. Discrete wavelet Mel Frequency Cepstral Coefficient (DMFCC) deals with MFCC analysis on the high-frequency components of speech rather than the low-frequency components. [28] explored GMFCC and DMFCC features on the Hindi language database.

Along with MFCC features, [29] presented MFCC, spectral centroid, crest, entropy, flatness, flux, skewness, and slope features for Tamil language emotion recognition.

The second phase includes feature classification using various classifiers. A variety of supervised and unsupervised machine learning models have been employed for this purpose. Hidden Markov model (HMM), support vector machine (SVM), Gaussian mixture model (GMM), K-nearest neighbor (KNN), artificial neural network (ANN), decision tree (DT), and random forest algorithms are some of them. In recent years, along with the traditional classification methods, several deep learning techniques are also being utilized for the classification process and have shown promising results. Convolutional neural network (CNN), long short-term memory (LSTM), deep CNN, and recurrent neural network (RNN) are the commonly used ones [10].

In [23] and [20], the LMM model, evaluation, and neural network classifier were explored using prosody and spectral features on various emotions such as anger, sadness, and fear. [9] explored auto-associative neural networks and SVM for Hindi language dialect identification and emotion recognition.

[27] utilized an HMM model for emotion classification of the Bangla and English languages. The authors compared the performance of HMM classifiers with other classifiers like GMM, KNN, ANN, and SVM on the SUBESCO-Bangla language database.[30] explored random forest and meta-iterative algorithms for Arabic, English, and Urdu language emotion classifications .

Lastly, based on feature classification, different emotions are recognized, such as anger, fear, sadness, sensory pleasure, amusement, satisfaction, contentment, excitement, disgust, contempt, pride, shame, guilt, embarrassment, and relief.

Although several studies have explored MFCC-based and multidomain feature representations for speech emotion recognition, relatively fewer works have investigated optimization-driven feature selection frameworks for multilingual SER. Existing studies primarily focus on improving classifier architectures and deep learning models, while the problem of redundant and high-dimensional handcrafted feature spaces remains comparatively less explored. Furthermore, optimization-based SER approaches involving metaheuristic algorithms such as Binary Grey Wolf Optimization (BGWO) for multilingual emotional speech datasets are still limited in current literature. The majority of existing works also concentrate on single-language SER systems, thereby restricting cross-language generalization analysis. These limitations motivate the development of an optimization-based multilingual SER framework capable of selecting discriminative multidomain feature subsets while reducing computational complexity.

3. PROPOSED FRAMEWORK

This work proposes a multilingual Speech Emotion Recognition (SER) framework based on multidomain handcrafted feature extraction and Binary Grey Wolf Optimization (BGWO)-based feature selection. The proposed framework is designed to improve emotional discriminative capability while reducing feature redundancy and computational complexity. The overall framework consists of three sequential phases: (i) baseline MFCC-based SER, (ii) multidomain feature-based SER, and (iii) BGWO-based feature optimization and classification. Bengali, Hindi, and Telugu emotional speech datasets are utilized for experimental evaluation.

3.1 Overall System Architecture

The overall architecture of the proposed multilingual SER framework consists of multiple stages including speech acquisition, preprocessing, feature extraction, feature optimization, and emotion classification. Initially, emotional speech samples from Bengali, Hindi, and Telugu datasets are collected and preprocessed. The preprocessing stage includes audio loading, normalization, and resampling to a uniform sampling frequency of 16 kHz to ensure consistency across datasets.

In the first phase, MFCC and Delta features are extracted to establish a baseline SER framework. In the second phase, multidomain handcrafted features are extracted from time-domain, spectral, cepstral, and time–frequency representations of speech signals. These features provide complementary emotional information for improved speech characterization. In the final phase, Binary Grey Wolf Optimization is employed to identify optimal feature subsets by eliminating redundant and less informative features. The optimized feature subsets are subsequently classified using machine learning classifiers to perform multilingual emotion recognition.

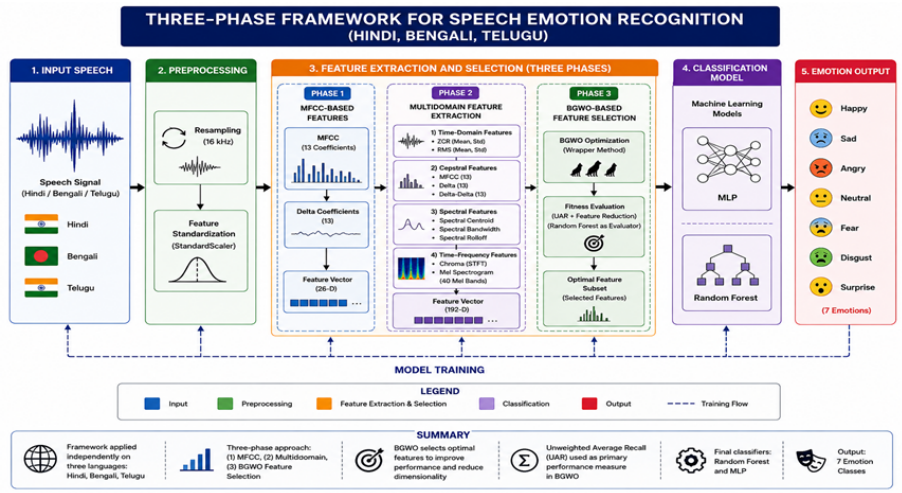

Fig. 1. Proposed Framework for Multilingual Indian Language Speech Emotion Recognition

The proposed framework shown in figure 1 is designed to analyze the effectiveness of multidomain feature engineering and optimization-driven feature selection for multilingual SER performance enhancement.

3.2 Phase 1 – MFCC-Based Speech Emotion Recognition

The first phase of the proposed framework focuses on developing a baseline SER system using Mel Frequency Cepstral Coefficients (MFCCs) and Delta features. Mel-Frequency Cepstral Coefficients (MFCCs) were extracted to capture the spectral characteristics of emotional speech signals. MFCCs model human auditory perception by mapping frequencies onto the Mel scale and have been extensively used in speech and emotion recognition applications [34].

MFCCs are widely used in speech processing because they effectively represent perceptual characteristics of human auditory perception. In this work, 13 MFCC coefficients are extracted from each speech sample. Additionally, 13 Delta coefficients are computed from the MFCC representations to capture temporal variations in speech dynamics. The combined MFCC and Delta representation results in a 26-dimensional feature vector for each speech sample.

After feature extraction, the feature vectors are normalized using standard scaling to improve classifier stability and convergence. The normalized features are then divided into training and testing subsets using stratified train–test splitting. Two machine learning classifiers, namely Multi-Layer Perceptron (MLP) and Random Forest (RF), are utilized for baseline emotion classification. The purpose of this phase is to establish a reference performance level for subsequent multidomain and optimization-based SER frameworks. Figure 2 represents flowchart for MFCC based SER system.

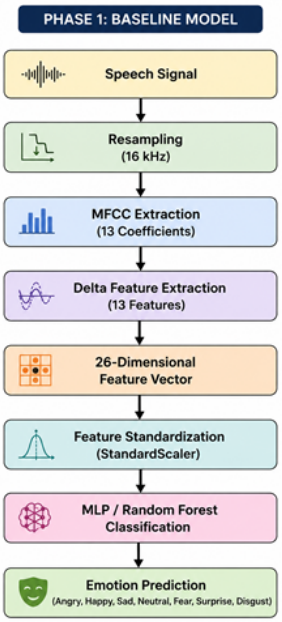

Fig. 2. MFCC based SER system

3.3 Phase 2 – Multidomain Feature-Based SER Framework

The second phase introduces a multidomain handcrafted feature extraction framework to improve emotional discriminative capability. Unlike MFCC-only representations, multidomain features capture complementary acoustic characteristics from multiple speech domains. Table 3.1 shows details of Multidomain features extracted .

The extracted feature set consists of time-domain, spectral, cepstral, and time–frequency features.The time-domain feature set includes Zero Crossing Rate (ZCR) and Root Mean Square (RMS) energy, which represent signal activity and energy variations. Spectral features such as spectral centroid, spectral bandwidth, and spectral roll-off are extracted to characterize frequency distribution properties of emotional speech signals. Cepstral features include MFCC, Delta MFCC, and Delta-Delta MFCC coefficients, which capture perceptual and temporal speech characteristics. Time–frequency representations including chroma features and Mel spectrogram features are also extracted to capture harmonic and spectral energy variations associated with emotions. Figure 3 represents a multidomain feature extraction system.

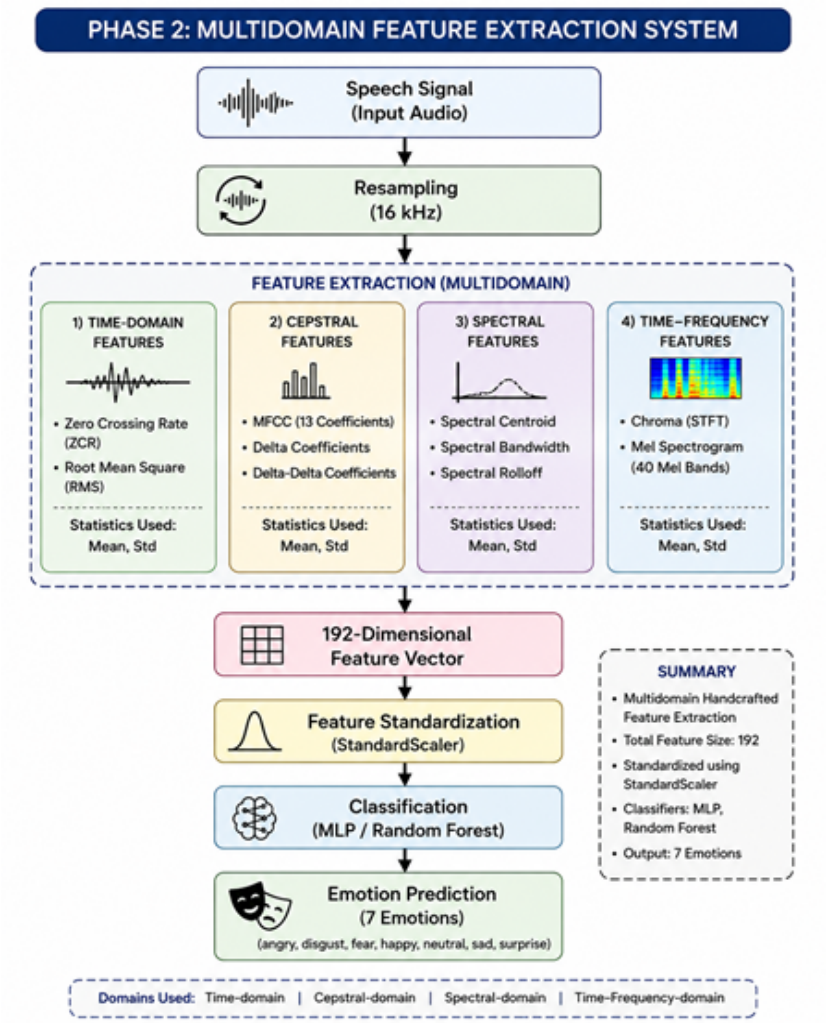

Fig. 3. Multidomain based SER system

For each extracted feature representation, statistical descriptors including mean and standard deviation are computed to generate fixed-length feature vectors. The multidomain framework produces a 192-dimensional handcrafted feature representation for each speech sample. These features are subsequently normalized and classified using MLP and RF classifiers to evaluate multilingual SER performance improvements over the MFCC baseline framework.

3.4 Phase 3 – BGWO-Based Feature Optimization Framework

Although multidomain handcrafted features improve SER performance, the resulting high-dimensional feature space may contain redundant and less informative features. To address this issue, the third phase of the proposed framework introduces Binary Grey Wolf Optimization (BGWO)-based feature selection for dimensionality reduction and feature optimization.[30].

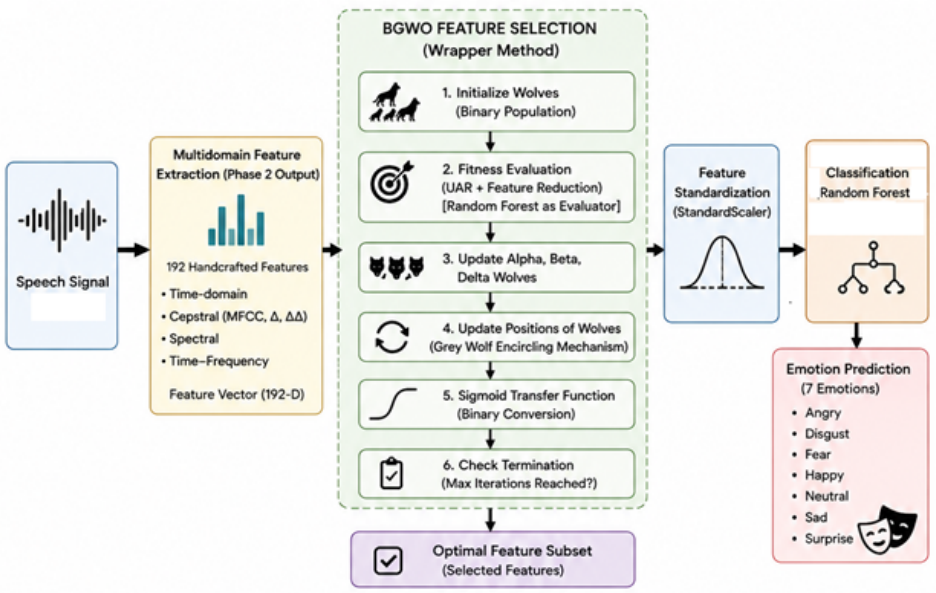

Fig. 4. BGWO-Based Feature Optimization Framework for Multilingual SER

BGWO is a population-based metaheuristic optimization algorithm inspired by the social hierarchy and hunting behavior of grey wolves. In the proposed framework mentioned in figure 4 , each search agent represents a binary feature selection vector where a value of '1' indicates feature selection and '0' indicates feature exclusion. The optimization process is guided by Alpha, Beta, and Delta wolves representing the best candidate solutions obtained during iterative search.

The fitness of each candidate feature subset is evaluated using a Random Forest classifier with macro-average recall-based performance evaluation. The optimization objective is to maximize classification performance while simultaneously minimizing the number of selected features. A sigmoid transfer function is employed to convert continuous position updates into binary feature selection decisions. Figure.5 represents the BGWO algorithm flowchart.

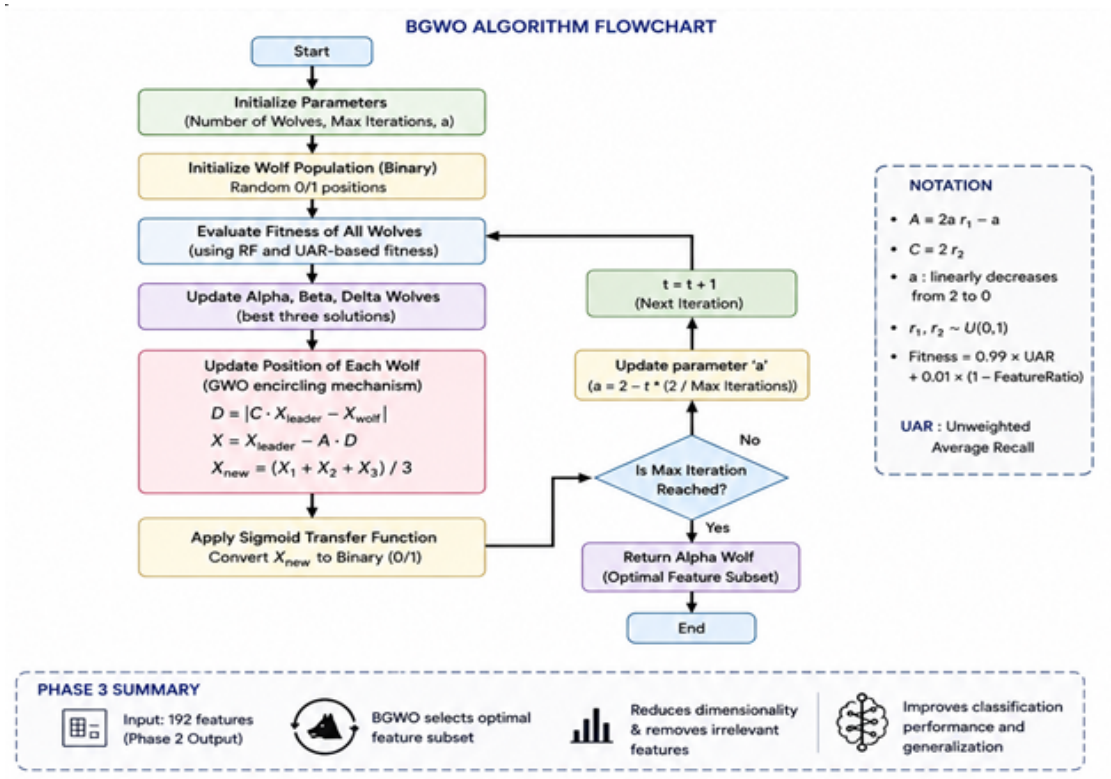

Fig. 5. Flowchart of the Proposed BGWO Optimization Process

Using the proposed BGWO framework, the original 192-dimensional multidomain feature space is reduced to optimized feature subsets for Bengali, Hindi, and Telugu SER datasets. The optimized features are subsequently classified using Random Forest classification to evaluate the effectiveness of optimization-based feature selection in multilingual SER systems. The proposed BGWO framework aims to improve feature efficiency, reduce redundancy, and maintain competitive emotion recognition performance across multilingual datasets.

4. MATHEMATICAL FORMULATION

This section presents the mathematical representation of the proposed multilingual Speech Emotion Recognition (SER) framework. The mathematical formulations describe the feature extraction process, optimization strategy, and feature selection mechanism employed in the proposed system. The MFCC extraction process follows the methodology proposed by Davis and Mermelstein [34]. Initially, statistical and acoustic feature extraction equations are utilized to characterize emotional speech signals from multiple feature domains. Subsequently, Binary Grey Wolf Optimization (BGWO) is mathematically modeled for optimal feature subset selection and dimensionality reduction. The presented formulations provide a theoretical foundation for the proposed SER framework and explain the computational procedures involved in multilingual emotion recognition.[31]

4.1 Feature Extraction Equations

The proposed multidomain SER framework utilizes several acoustic and statistical features extracted from emotional speech signals. These features represent complementary characteristics from different speech domains.

Zero Crossing Rate (ZCR)

Zero Crossing Rate measures the rate at which the speech signal changes sign within a given frame. It is useful for analyzing speech activity and signal variability.

$$ZCR = \frac{\left(\frac{1}{T-1}\right)\sum_{t=1}^{T-1}\left|sign(x(t)) - sign(x(t-1))\right|}{2} \quad (1)$$

where:

- x(t) represents the speech signal sample at time t
- T denotes the total number of samples in the frame

A higher ZCR generally indicates rapid signal fluctuations, which may correspond to specific emotional speech characteristics.

Root Mean Square Energy (RMS)

Root Mean Square energy measures the average signal energy within a speech frame and reflects emotional intensity variations in speech.

Where:

$$RMS = \sqrt{(1/T \cdot \Sigma\, x(t)^2)} \quad (2)$$

- x(t) denotes the speech signal amplitude
- T represents the frame length

RMS energy is particularly useful for distinguishing emotions associated with high or low speech intensity.

4.2 BGWO Fitness Function

The proposed BGWO framework performs feature subset optimization by maximizing classification performance while simultaneously minimizing feature dimensionality. The feature selection ratio is computed as:

$$Feature\_Ratio = \frac{\{|selected\_features|\}}{\{192\}} \quad (3)$$

where:

- |selected_features| represents the number of selected features
- 192 denotes the original multidomain feature dimension

The overall fitness function is defined as:

$$Fitness = 0.99 \times UAR + 0.01 \times (1 - Feature\_Ratio) \quad (4)$$

where:

- UAR represents Unweighted Average Recall
- Feature_Ratio denotes the proportion of selected features

The fitness function prioritizes classification performance while introducing a small penalty for higher feature dimensionality.

4.3 BGWO Hunting and Position Update Mechanism

Binary Grey Wolf Optimization simulates the hunting behavior and leadership hierarchy of grey wolves. The optimization process is guided by Alpha, Beta, and Delta wolves representing the best candidate solutions.[31].The hunting mechanism, leadership hierarchy, and position update equations follow the original GWO formulation presented in [31], whereas the binary feature selection strategy is based on the BGWO approach described in [31].

The distance between the search agent and Alpha wolf is computed as:

$$D_{\{\alpha\}} = \left| C_1 \cdot X_{\{\alpha\}} - X \right| \quad (5)$$

The updated position corresponding to the Alpha wolf is given by:

$$X_1 = X_{\{\alpha\}} - A_1 \cdot D_{\{\alpha\}} \quad (6)$$

$$D_{\{\beta\}} = \left| C_2 \cdot X_{\{\beta\}} - X \right| \quad (7)$$

Similarly, the distance and updated positions for Beta and Delta wolves are computed as:

$$X_2 = X_{\{\beta\}} - A_2 \cdot D_{\{\beta\}} \quad (8)$$

$$D_{\{\delta\}} = \left| C_3 \cdot X_{\{\delta\}} - X \right| \quad (9)$$

The final position update equation is expressed as:

$$X(t+1) = \frac{\{X_1 + X_2 + X_3\}}{\{3\}} \quad (10)$$

where:

- Xα, Xβ, and Xδ represent the positions of Alpha, Beta, and Delta wolves
- A and C denote coefficient vectors controlling exploration and exploitation
- X(t+1) represents the updated search position

This averaging mechanism enables balanced exploration and exploitation during feature subset optimization.

4.4 Sigmoid Transfer Function

Since BGWO operates in a binary search space for feature selection, a sigmoid transfer function is utilized to convert continuous position updates into binary decisions.[31].

$$S(x) = \frac{1}{1 + e^{-x}} \quad (11)$$

The sigmoid function maps continuous position values into probabilities between 0 and 1, enabling binary feature selection during optimization. Features corresponding to higher sigmoid probabilities are selected, while lower probability features are discarded during the optimization process.

5. EXPERIMENTAL SETUP

This section describes the experimental configuration used to evaluate the proposed multilingual Speech Emotion Recognition (SER) framework. The experimental setup includes dataset description, preprocessing configuration, feature extraction settings, classifier parameters, BGWO optimization settings, and evaluation metrics used for performance analysis.

5.1 Emotional Speech Datasets

The proposed multilingual Speech Emotion Recognition (SER) framework is evaluated using emotional speech datasets from three Indian languages: Bengali, Hindi, and Telugu. The datasets consist of professionally recorded acted emotional speech samples representing multiple emotional categories. These datasets were selected to analyze the effectiveness of the proposed multidomain feature optimization framework across linguistically diverse Indian languages. Table1 shows a summary of three datasets used for this study.

Table 1. Summary of three datasets

Dataset	Language	Speakers	Utterances Used	Emotions	Sampling Rate
SUBESCO	Bengali	20(10Male+10 Female)	7000	Angry, Disgust, Fear, Happy, Neutral, Sadness, Surprise	16 kHz
IITKGP-SEHSC	Hindi	10(5Male + 5 Female)	7026	Angry, Disgust, Fear, Happy, Neutral, Sadness, Sarcastic, Surprise	16 kHz
IITKGP-SESC	Telugu	10 (5 Male + 5 Female)	800	Angry,Compassion, Disgust, Fear, Happy, Neutral, Sarcastic, Surprise	16 kHz

The Bengali emotional speech dataset used in this work is the SUBESCO (Sentiment and Emotion Labelled Bengali Speech Corpus), which is a publicly available acted emotional speech corpus developed for Bengali SER research. The corpus contains 7000 emotional utterances recorded from 20 speakers consisting of equal male and female participation. The recordings are provided in .wav format and were originally recorded at 48 kHz before being resampled to 16 kHz for experimental consistency. The dataset includes seven emotional categories: Angry, disgust, fear, happiness, neutral, sadness, and surprise.

The Hindi dataset utilized in this work is the IITKGP-SEHSC (IIT Kharagpur Simulated Emotion Hindi Speech Corpus). The dataset consists of professionally acted emotional speech recordings collected from artists associated with All India Radio (AIR) Varanasi. A total of 7026 emotional utterances are utilized for

experimentation. The corpus includes eight emotional categories: Angry, disgust, fear, happy, neutral, sadness, sarcastic, and surprise.

Similarly, the Telugu dataset employed in this work is the IITKGP-SESC (IIT Kharagpur Simulated Emotion Speech Corpus). The dataset contains emotional speech recordings from professional artists associated with AIR Vijayawada. A total of 800 utterances are considered for experimentation. The emotional categories included in the dataset are Angry, compassion, disgust, fear, happy, neutral, sarcastic, and surprise.

Since the emotional categories are not identical across all three datasets, multilingual comparative analysis is performed using the common emotional categories shared among all languages. The common emotions considered for comparative evaluation are:

- Angry
- Disgust
- Fear
- Happy
- Neutral
- Surprise

This common-emotion evaluation strategy ensures fair multilingual comparison and enables consistent analysis of language-dependent emotional speech characteristics. The multilingual dataset configuration further supports the evaluation of the robustness and generalization capability of the proposed BGWO-based multidomain SER framework across different Indian languages.

5.2 Speech Preprocessing Configuration

All speech recordings are preprocessed before feature extraction to ensure consistency across multilingual datasets. Audio signals are loaded using the Librosa library and resampled to a sampling frequency of 16 kHz. Feature normalization is performed using standard scaling to transform features into zero mean and unit variance distributions. This normalization process improves classifier stability and convergence during training.

The datasets are divided into training and testing subsets using stratified train–test splitting with an 80:20 ratio to preserve class distribution across emotional categories.

5.3 Phase 1 Experimental Configuration

In Phase 1, baseline SER experiments are conducted using MFCC and Delta feature representations. Initially, 13 MFCC coefficients are extracted from each speech signal. Subsequently, 13 Delta coefficients are computed from MFCC representations to capture temporal speech dynamics. The combined feature representation results in a 26-dimensional feature vector.

Two machine learning classifiers are utilized for baseline evaluation:

- Multi-Layer Perceptron (MLP)
- Random Forest (RF)

The Multilayer Perceptron classifier was configured with two hidden layers of 256 and 128 neurons respectively, ReLU activation function, Adam optimizer, adaptive learning rate, a batch size of 32, and a maximum of 300 training iterations.The Random Forest classifier was configured with 300 decision trees (n_estimators=300) with parallel processing enabled (n_jobs=-1).

5.4 Phase 2 Experimental Configuration

In Phase 2, a comprehensive 192-dimensional multidomain feature vector was extracted from each utterance spanning time-domain, frequency-domain, and time-frequency domain representations. Table2. presents the complete feature extraction configuration used in Phase 2.

Table2 . Phase 2 — Multidomain Feature Description

Domain	Features	Statistics	Count
Time Domain	ZCR, RMS Energy	Mean, Std	4
MFCC Family	MFCC, Delta MFCC, Delta-Delta MFCC (13 each)	Mean, Std	78
Spectral	Spectral Centroid, Bandwidth, Rolloff	Mean, Std	6
Chroma	Chroma STFT (12 bins)	Mean, Std	24
Mel Spectrogram	Mel Spectrogram dB (40 bands)	Mean, Std	80
Total			192

For each feature representation, statistical descriptors including mean and standard deviation are computed. The resulting multidomain representation produces a 192-dimensional handcrafted feature vector for each speech sample.

The extracted multidomain features are evaluated using:

- Multi-Layer Perceptron (MLP)
- Random Forest (RF)

The MLP architecture consists of three hidden layers with ReLU activation and dropout regularization to reduce overfitting. The output layer uses Softmax activation for multiclass emotion classification.

5.5 Phase3 - BGWO Optimization Configuration

In Phase 3, BGWO was employed to identify an optimal subset of multidomain acoustic features while minimizing feature redundancy and maximizing classification performance.[30]. The Binary Grey Wolf Optimization (BGWO) algorithm is a wrapper-based feature selection method applied on the 192-dimensional feature set from Phase 2. The BGWO algorithm was configured with a population of 12 wolves and executed for 25 iterations. The sigmoid transfer function was used to map continuous position values to binary decisions for feature inclusion or exclusion. The Random Forest classifier with 100 trees and 3-fold Stratified K-Fold cross validation served as the fitness function within the BGWO wrapper framework. The fitness function is defined as: Fitness = 0.99 × UAR + 0.01 × (1 − Feature Ratio) where UAR denotes the Unweighted Average Recall and Feature Ratio represents the proportion of selected features to total features. The coefficient 0.99 assigned to UAR reflects the primary emphasis on classification performance, while the penalty term 0.01 × (1 − Feature Ratio) encourages dimensionality reduction. An empty feature subset was assigned a fitness value of zero to prevent degenerate solutions. Following BGWO based feature selection, the final Random Forest classifier with 300 trees was trained and evaluated on the selected feature subsets using the same 80/20 stratified train-test split. The BGWO algorithm is configured as shown in Table3 .

Table 3. BGWO Configuration parameters

Parameter	Value
Population size (wolves)	12
Maximum iterations	25
Transfer function	Sigmoid
Fitness function	UAR based RF wrapper
Cross validation	3-fold Stratified KFold
Alpha weight (UAR)	0.99
Feature penalty weight	0.01
Final classifier	Random Forest(300 trees)
Random seed	42

The fitness function simultaneously maximizes classification performance and minimizes feature dimensionality. Random Forest classification is utilized during optimization to evaluate candidate feature subsets.

5.6 Classifier Configuration

Random Forest (RF), proposed by Breiman [32], is an ensemble learning classifier that constructs multiple decision trees and combines their outputs through majority voting to improve classification performance. The Random Forest classifier is configured using 300 estimators with parallel processing enabled for efficient training. The classifier uses square-root-based feature selection during tree construction and employs ensemble averaging for robust emotion classification.

Multi-Layer Perceptron (MLP) is a feed-forward artificial neural network trained using the backpropagation learning algorithm proposed by Rumelhart et al. [33]. The Multi-Layer Perceptron classifier utilizes:

- Hidden layers: (256, 128, 64)
- Activation function: ReLU
- Optimizer: Adam
- Batch size: 32
- Epochs: 40
- Loss function: Categorical Crossentropy

Dropout regularization is applied between hidden layers to improve model generalization performance.

5.7 Evaluation Metrics

The proposed multilingual SER framework is evaluated using multiple performance metrics including:

- Accuracy
- Precision
- Recall
- F1-Score
- Unweighted Average Recall (UAR)
- Confusion Matrix Analysis

Accuracy measures overall classification performance, while precision, recall, and F1-score provide class-wise performance analysis. UAR is utilized during BGWO optimization because it provides balanced evaluation across emotional categories, particularly for class-imbalanced datasets. Emotion-wise results are also analysed by confusion matrices to examine behavior and inter-emotion confusion patterns across multilingual datasets.

6. RESULTS AND DISCUSSION

This section presents the experimental results obtained from the proposed multilingual Speech Emotion Recognition (SER) framework across Bengali, Hindi, and Telugu emotional speech datasets. The performance of the proposed framework is analyzed in three phases: baseline MFCC-based SER, multidomain handcrafted feature-based SER, and BGWO-based feature optimization. The obtained results are evaluated using accuracy, precision, recall, F1-score, and confusion matrix analysis. Comparative analysis is further performed to examine the effectiveness of multidomain feature extraction and optimization-based feature selection across multilingual emotional speech datasets.

6.1 Phase 1: MFCC Based Speech Emotion Recognition

As described in Section 3.2 , Phase 1 employed a 26-dimensional feature vector comprising 13 MFCC coefficients and their first-order delta derivatives extracted from each speech utterance. MLP and RF classifiers were independently trained and evaluated on each of the three language datasets — Bengali, Hindi, and Telugu — using an 80/20 stratified train-test split. The classification performance for each language is presented in the following subsections.

A consolidated summary of Phase 1 classification performance across all three language datasets is presented in Table 4.

.

Table 4. Phase 1 Summary — Classification Performance Across Languages

Language	Emotions	MLP Accuracy	Precision/ Recall /F1-score for MLP classifier	RF Accuracy	Precision/Recall/F1-Score for RF Clas
Bengali	7	75%	0.75	72%	0.72
Hindi	8	60%	0.60	56%	0.56
Telugu	8	55%	0.56	56%	0.56

The Phase 1 results establish a baseline performance using MFCC-26 features across three Indian languages. Bengali achieved the highest recognition accuracy of 75% , attributed to its larger and more balanced dataset of 7000 utterances. Hindi and Telugu showed comparatively lower accuracy due to the presence of 8 emotion categories including subjective emotions such as Sarcastic and Compassion that are acoustically less distinct in the MFCC feature space. The consistently lower performance of Disgust and Sarcastic across all languages suggests overlapping acoustic characteristics, motivating the need for a richer multidomain feature representation explored in Phase 2.

6.1 .1 Emotion-wise Results for Phase 1

The emotion wise recognition results for MLP and RF classifiers for phase1 are presented in Table For the Bengali dataset in Phase 1. The MLP confusion matrix reveals that Neutral (84.5%) and Angry (83.5%) were the best recognized emotions, while Disgust recorded the lowest diagonal accuracy of 62%. Fear and Happy showed moderate recognition of 81% and 74% respectively. The RF confusion matrix demonstrates a similar pattern with Neutral achieving the highest diagonal accuracy of 90%, followed by Happy (79%) and Angry (80.5%). However Disgust showed a significant drop to 48.5% indicating that RF struggled more with Disgust discrimination compared to MLP. Overall the MLP demonstrated better emotion discrimination for Bengali with fewer misclassification instances across most emotion categories.

Table 5. Phase1 Emotion Recognition results for Bengali , Hindi and Telugu a). MLP b). RF Classifiers

a.	MFCC- MLP Results			b. MFCC-RF Results			
Emotion	Bengali %	Hindi %	Telugu %	**Emotion**	Bengali %	Hindi %	Telugu %
Angry	83.5	72.88	40	Angry	80.5	77.4	45
Disgust	62	50.28	45	Disgust	48. 5	49.15	50
Fear	81	71.51	55	Fear	77.5	70.95	65
Happy	74	52.57	80	Happy	79	46.29	55
Neutral	84.5	68.79	40	Neutral	90	56.65	60
Surprise	67	58.05	60	Surprise	69.5	49.23	45

For Hindi language SER, The MLP confusion matrix shows that Angry (72.88%) and Fear (71.51%) are the best recognized emotions. Disgust also showed poor recognition at 50.28% The RF confusion matrix reveals a similar trend with Angry achieving the highest recognition at 77.4%. Disgust (49.15%) remained the most challenging emotions, suggesting that the 26-dimensional MFCC feature space was insufficient to capture the subtle acoustic differences between low arousal emotions in Hindi.

For emotion -wise Recognition in Telugu language , The MLP confusion matrix shows that Happy (80%) is the best recognized emotions, while Angry (40%) and Neutral (40%) recorded the lowest diagonal accuracies.

The RF demonstrates improved recognition for Fear (65%) compared to MLP. However Angry (45%) and Surprise (45%) remained poorly recognized. The overall lower performance in Telugu across both classifiers can be attributed to the limited dataset size of 800 utterances, resulting in only 20 test samples per emotion class which increases sensitivity to misclassification.

6.2 Phase 2: Multidomain Feature Based Speech Emotion Recognition

Phase 2 extended the feature representation from the 26-dimensional MFCC baseline to a comprehensive 192-dimensional multidomain handcrafted feature vector encompassing time-domain, frequency-domain, spectral, chroma, and Mel-based descriptors as described in Section 3.4 The richer feature space aimed to capture a wider range of acoustic characteristics of emotional speech across the three Indian language datasets. MLP and RF classifiers were independently trained and evaluated for each language using the same 80/20 stratified train-test split. The classification performance and confusion matrix analysis for each language are presented in the following subsections. A consolidated summary of Phase 2 classification performance across all three language datasets along with improvement over Phase 1 is presented in Table 6.

Table 6. Phase 2 Summary — Classification Performance Across Languages

Language	MLP Accuracy	Precision/Recall/F1-score for MLP Classifier	RF Accuracy	Precision/Recall/F1-Scorefor RF Classifier
Bengali	75%	0.75	72%	0.72
Hindi	60%	0.56	56%	0.6
Telugu	55%	0.55	56%	0.57

Phase 2 results confirm that the 192-dimensional multidomain feature representation significantly outperforms the 26-dimensional MFCC baseline across all three Indian language datasets. Bengali achieved the highest overall accuracy of 81% with RF, benefiting from the balanced and larger dataset. Telugu demonstrated the most dramatic improvement of 23% for MLP, highlighting the effectiveness of multidomain features for smaller datasets. Hindi MLP showed an 21% improvement, though Hindi RF showed only marginal gains of 2%, suggesting that the RF ensemble approach was less effective for the complex 8-emotion Hindi recognition task. Despite these improvements, the high dimensionality of 192 features introduced computational complexity and potential overfitting risks, motivating the application of BGWO based feature selection in Phase 3 to identify the most discriminative feature subset while maintaining recognition accuracy.

6.2.1 Emotion-Wise Results for Phase 2

The emotion-wise results analysis for MLP and RF classifiers respectively for the all the languages are shown in Table 7.

Table 7. Phase2 Emotion Recognition results for Bengali , Hindi and Telugu a). MLP b). RF Classifiers

a. Phase2- MLP				b. Phase2 -RF			
Emotion	Bengali %	Hindi %	Telugu %	Emotion	Bengali %	Hindi %	Telugu %
Angry	91.5	85.3	60	Angry	88.5	78.53	85
Disgust	77	71.8	75	Disgust	67	61.02	75
Fear	90.5	83.8	80	Fear	83	73.74	80
Happy	83	81.7	85	Happy	81.5	63.43	70
Neutral	93	80.3	85	Neutral	91.5	71.1	90
Surprise	82	85.4	90	Surprise	83	66.67	80

The MLP emotion recognition results reveal dramatic improvement in recognition accuracy across all emotions compared to Phase 1. Neutral achieved the highest diagonal accuracy of 93%, followed by Angry (91.5%) and Fear (90.5%), reflecting the effectiveness of multidomain features in capturing distinct acoustic patterns for these emotions. Disgust showed improvement from 62% to 77% though it remained the most challenging emotion .

The RF emotion recognition results for Bengali similarly demonstrated strong improvement with Neutral (91.5%) and Angry (88.5%) being the best recognized emotions. Disgust dropped to 67% for RF compared to 77% for MLP.

The emotion-wise recognition results analysis for Hindi Phase 2 shows that Surprise(83.6%) and Angry (83.3%) achieved the highest recognition accuracy . Disgust remained the most challenging emotion at 71.8% Happy showed 81.7 % diagonal accuracy indicating less overlap .

With the RF classifier Angry achieved the highest diagonal accuracy of 78.5% for RF classifiers, followed by Fear (73.7%) and neutral (71.1%) . Happy showed 63.4 % diagonal accuracy indicating less overlap .

As shown in Table 7, The Telugu results showed significant improvement across all emotion categories compared to Phase 1. Happy (85%), Surprise (90%), and Neutral (85%) achieved the highest diagonal accuracies,

while Disgust (75%) and Angry (60%) remained the most challenging emotions.

The RF emotion -wise analysis demonstrated strong recognition for Neutral (90%), Angry (85%), and Sarcastic (85%), showing complementary strengths compared to MLP. Surprise achieved 80% diagonal accuracy for RF .

.3 Phase 3: Binary Gray Wolf Optimization (BGWO)

Phase 3 applied the Binary Grey Wolf Optimization (BGWO) algorithm as a wrapper based feature selection method on the 192-dimensional multidomain feature set obtained from Phase 2. As described in Section 3.4 the Random Forest classifier with 3-fold Stratified K-Fold cross validation served as the fitness function within the BGWO framework, with the fitness value computed as a weighted combination of Unweighted Average Recall (UAR) and feature reduction ratio. The primary objective of Phase 3 was to address the curse of dimensionality introduced by the 192-dimensional feature space while maintaining or improving classification accuracy. The feature reduction achieved by BGWO across all three language datasets is presented in Table 8.

Table 8. BGWO Feature Selection Results

Language	Phase-2 Features	Selected Features	Reduction %	Phase 2 RF Accuracy	Phase 3 BGWO Accuracy	Change
Bengali	192	130	32%	81%	81%	0%
Hindi	192	123	36%	58%	68%	10%
Telugu	192	133	31%	72%	66%	-6%

The BGWO algorithm successfully reduced the feature dimensionality by 31-36% across all three language datasets, identifying the most discriminative feature subsets for each language independently. The language specific feature selection reflects the varying acoustic characteristics of Bengali, Hindi, and Telugu emotional speech, with Hindi requiring the most aggressive feature reduction (36%) and Telugu the least (31%). The classification results using BGWO selected features are presented in the following subsections.

It is important to note that the Random Forest classifier served dual roles in Phase 3. First, a lightweight RF with 100 trees was employed as the fitness function within the BGWO wrapper framework to evaluate candidate feature subsets during optimization using 3-fold Stratified K-Fold cross validation. Second, upon completion of BGWO optimization, a more robust RF with 300 trees was trained exclusively on the BGWO selected feature subset to produce the final classification results reported in this section.

The Phase 3 results demonstrate the effectiveness of BGWO based feature selection in reducing feature dimensionality across all three Indian language datasets. The BGWO algorithm successfully reduced the 192-dimensional feature space by 31-36% across all languages while maintaining or improving classification accuracy for two out of three languages. Bengali successfully maintained its peak accuracy of 81% with a 32% feature reduction, confirming that a significant portion of the 192-dimensional feature space was redundant for Bengali emotion recognition. Hindi demonstrated the most significant improvement of 10% accuracy gain alongside a 36% feature reduction, establishing BGWO as a highly effective optimization strategy for Hindi SER by eliminating noisy and redundant features that were degrading RF performance. Telugu showed a marginal accuracy decrease of 6% which is attributed to the limited dataset size of 800 utterances constraining the BGWO optimization process. The convergence curves confirmed that BGWO successfully converged within the 25 iteration limit for all three languages, validating the adequacy of the selected optimization parameters."

6.3.1 BGWO Convergence Analysis

Figure 6.1 illustrates the BGWO convergence curves for Bengali, Hindi and Telugu datasets respectively, depicting the progression of the best fitness value across 25 iterations. The fitness value represents the weighted combination of Unweighted Average Recall (UAR) and feature reduction ratio as defined in Section 5.5 . A consistently increasing or stable convergence curve confirms that the BGWO algorithm successfully identified progressively better feature subsets across iterations without divergence. All three convergence curves demonstrate monotonically non-decreasing fitness values confirming the correctness of the BGWO implementation.The Bengali convergence curve demonstrates a distinctive stepwise improvement pattern across 25 iterations.

The fitness value remained stable at approximately 0.7703 during the initial iterations 1 through 9, indicating that the wolf population was exploring the feature space without finding a significantly better solution. A notable jump occurred at iteration 10 where the fitness value rose sharply to approximately 0.7725, indicating that the alpha wolf discovered a substantially better feature subset. The fitness value then stabilized again from iterations 11 through 22 as the wolf population converged around this improved solution. A second significant improvement was observed at iteration 23 where the fitness value rose further to 0.7734 representing the final optimal feature subset. The curve remained stable at this value through iteration 25 confirming successful convergence. This stepwise pattern is characteristic of binary optimization algorithms where discrete feature inclusion or exclusion decisions lead to sudden improvements in the fitness landscape rather than gradual continuous

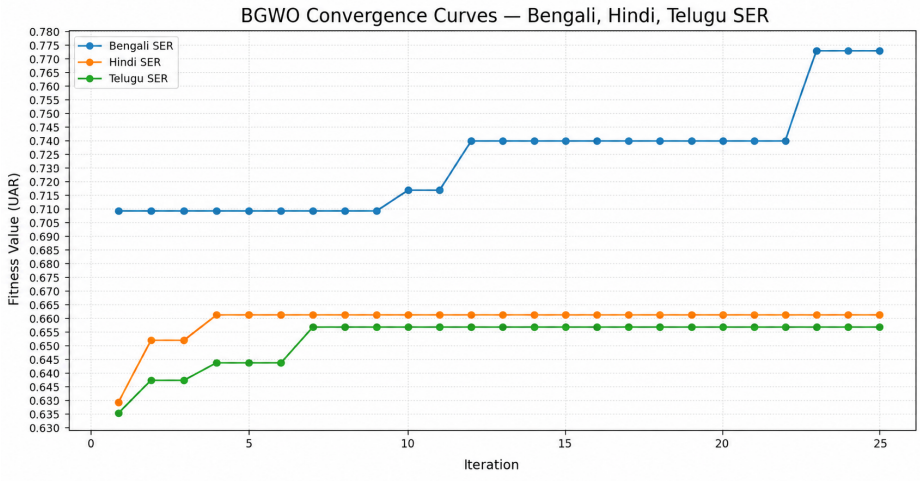

Fig. 6. BGWO Convergence Curve — Bengali, Hindi and Telugu SER

The Hindi convergence curve exhibits the most rapid convergence among the three languages demonstrating a sharp improvement pattern in the early iterations. The fitness value started at approximately 0.638 at iteration 1 reflecting the initial random wolf population. A steep rise was observed between iterations 1 and 5 where the fitness value jumped dramatically to approximately 0.665 indicating that the BGWO algorithm quickly identified the most discriminative feature subset for Hindi in the very early stages of optimization. From iteration 5 onwards the curve remained completely stable at 0.665 for the remaining 20 iterations confirming early and definitive convergence. This rapid convergence suggests that the redundant and noisy features in the 192-dimensional Hindi feature space were clearly distinguishable from the informative ones enabling BGWO to efficiently identify the optimal subset within just 5 iterations.

The Telugu convergence curve demonstrates a multi-step improvement pattern with multiple small incremental jumps in the early iterations before reaching stable convergence. The fitness value started at approximately 0.635 at iteration 1 and showed a series of progressive improvements — rising to 0.649 by iteration 2, further to 0.653 by iteration 4, before making a significant jump to approximately 0.660 by iteration 7 where it stabilized for the remaining iterations. This gradual multi-step pattern reflects the more complex optimization landscape for Telugu, where the limited dataset size of 800 utterances creates a noisier fitness landscape requiring multiple incremental improvements before convergence. The curve remained stable from iteration 7 through iteration 25 confirming successful convergence well within the 25 iteration limit.

A comparative analysis of the three convergence curves reveals distinct optimization characteristics for each language. Hindi exhibited the fastest convergence at iteration 5 with the largest fitness improvement of 0.027, suggesting clear feature discriminability in the Hindi acoustic space. Telugu converged at iteration 7 with a fitness improvement of 0.025 through multiple incremental steps. Bengali showed the slowest convergence at iteration 23 with the smallest fitness improvement of 0.0031, which can be attributed to the larger dataset size of 7000 utterances creating a more complex optimization landscape. Notably Bengali achieved the highest absolute fitness value of 0.7734 compared to 0.665 for Hindi and 0.660 for Telugu, reflecting the superior classification performance achievable on the larger and more balanced Bengali dataset. All three curves confirm that the BGWO algorithm successfully converged within the 25 iteration limit validating the adequacy of the optimization parameters selected for this study.

6.3.2 Emotion-Wise Results for Phase 3- BGWO with RF

The emotion-wise results of BGWO with RF classifier for the Bengali, JHindi and Telugu language datasets are shown in Table 9

Table 9. Phase 3 Summary — Emotion-wise BGWO Results

Phase3=BGWO with RF			
Emotion	**Bengali %**	**Hindi %**	**Telugu %**
Angry	89.5	77.97	45
Disgust	66.5	58.19	60
Fear	85	75.42	80
Happy	84	60.57	65
Neutral	92.5	73.41	75
Surprise	83.5	67.82	80

These results analysis reveals that Neutral achieved the highest diagonal accuracy of 92.5%, followed by Angry (89.5%) and Fear (85%). Compared to Phase 2 RF, most emotions showed marginal improvements — Fear improved from 83% to 85%, Happy from 81.5% to 84%, and Neutral from 91.5% to 92.5%, confirming that the BGWO selected feature subset was more discriminative than the full 192-feature set. Overall BGWO achieved the dual objective of feature reduction and accuracy maintenance for Bengali.The Telugu dataset showed a slight decrease in accuracy from 72% in Phase 2 RF to 66% in Phase 3 BGWO, despite achieving a 31% feature reduction from 192 to 133 features. This marginal performance drop can be attributed to the limited dataset size of 800 utterances with only 20 test samples per emotion class, which makes the optimization landscape more sensitive to feature subset selection. Angry showed a notable drop from 85% to 45%. , suggesting that the BGWO selected feature subset was less effective for distinguishing Angry in the limited Telugu dataset. Neutral dropped from 90% to 75%. The limited dataset size of

Telugu with only 20 test samples per emotion class amplifies the impact of individual misclassifications, making it difficult for BGWO to find a consistently optimal feature subset across all emotion categories.

7. DISCUSSION

The experimental findings demonstrate that the choice of acoustic feature representation plays a critical role in multilingual Speech Emotion Recognition (SER). Comparative analysis across all three phases demonstrates the effectiveness of multidomain feature extraction and BGWO-based optimization for multilingual SER. The baseline MFCC-based framework used in Phase 1 provided moderate classification performance across Bengali, Hindi, and Telugu datasets; however, the limited 26-dimensional cepstral representation was insufficient for capturing the complex emotional dynamics present in multilingual speech signals. The emotion-wise results analysis revealed increased inter-class overlap among emotionally similar categories such as *Disgust* , *Fear, Surprise* indicating the limitations of relying solely on cepstral descriptors.

The multidomain feature framework introduced in Phase 2 substantially improved emotional representation capability by integrating complementary information from time-domain, spectral, cepstral, and time–frequency acoustic features. Using the 192-dimensional multidomain feature set, the Random Forest classifier achieved accuracies of 81% for Bengali, 58% for Hindi, and 72% for Telugu datasets. The improved performance confirms that emotional speech characteristics are distributed across multiple acoustic domains rather than a single feature space. The multidomain representation particularly enhanced the recognition of common emotions including *Anger*, *Happy*, *Neutral*, and *Surprise* across all three languages.

Among the evaluated datasets, Bengali demonstrated the most stable performance throughout all experimental phases. This behavior can be attributed to the comparatively larger dataset size, balanced emotional distribution, and greater number of utterances available in the SUBESCO corpus. In contrast, the Hindi and Telugu datasets exhibited comparatively higher inter-emotion confusion because of overlapping prosodic variations and differences in emotional expression patterns across speakers and languages.

The BGWO optimization framework introduced in Phase 3 further improves feature efficiency by eliminating redundant and less informative acoustic representations. Experimental analysis indicates that optimization-based feature selection can substantially reduce feature dimensionality while preserving or even improving classification performance in certain multilingual SER scenarios.The optimization process successfully reduced the feature dimensionality by approximately 32% for Bengali, 36% for Hindi, and 31% for Telugu datasets. Interestingly, Hindi achieved a 10% improvement in classification accuracy after optimization, increasing from 58% to 68%. This result indicates that the original multidomain feature space contained redundant or less informative features that negatively influenced classification performance.

For the Bengali dataset, BGWO maintained the same classification accuracy of 81% even after reducing the feature dimension from 192 to 130 features, demonstrating effective redundancy elimination without information loss. However, the Telugu dataset showed a slight reduction in performance after optimization, decreasing from 72% to 66%. This variation may be associated with the comparatively smaller dataset size and limited emotional utterances available for optimization and classifier training.

The convergence analysis additionally confirmed the stability of the BGWO algorithm during feature subset selection. The optimization process demonstrated rapid fitness improvement during the initial iterations followed by gradual convergence toward stable fitness values. The observed convergence behavior validates the effectiveness of BGWO in balancing exploration and exploitation while identifying discriminative acoustic feature subsets for multilingual SER.

Overall, the experimental results confirm that combining multidomain acoustic representations with optimization-based feature selection provides an effective framework for multilingual SER across Indian languages. The proposed approach not only improves emotional recognition capability but also reduces feature dimensionality and computational complexity, making it suitable for efficient multilingual emotional speech analysis systems.

8. CONCLUSION

This paper presented a multilingual Speech Emotion Recognition (SER) framework for Bengali, Hindi, and

Telugu languages using multidomain handcrafted acoustic features, machine learning classifiers, and Binary Grey Wolf Optimization (BGWO)-based feature selection. The study systematically evaluated the influence of feature representation and optimization on multilingual SER performance through three experimental phases. The initial MFCC-based baseline framework demonstrated that conventional cepstral features provide limited emotional representation capability for multilingual SER, particularly for emotionally overlapping categories. To address this limitation, a multidomain handcrafted feature framework consisting of time-domain, spectral, cepstral, and time–frequency acoustic descriptors was introduced. The multidomain representation significantly improved classification performance across all evaluated languages, confirming the effectiveness of integrating complementary acoustic information for emotional characterization. The proposed BGWO-based optimization framework further improved feature efficiency by identifying informative feature subsets while eliminating redundant acoustic representations. Experimental results demonstrated substantial dimensionality reduction of approximately 31–36% across multilingual datasets. Notably, the Hindi dataset achieved a significant improvement in classification accuracy following optimization, whereas the Bengali dataset maintained stable performance despite considerable feature reduction. These findings validate the capability of BGWO to improve feature compactness without substantially affecting emotional discriminative performance. The study additionally demonstrated that classifier behavior varies depending on feature representation and dataset characteristics. The MLP classifier showed strong learning capability during multidomain classification, whereas the Random Forest classifier proved effective as both a standalone classifier and a fitness evaluator during optimization. The emotion-wise matrix analysis further revealed that common emotions such as *Anger*, *Happy*, *Neutral*, and *Surprise* achieved comparatively better recognition consistency across languages, while emotionally similar categories continued to exhibit moderate inter-class confusion. Overall, the proposed framework establishes an effective and computationally efficient approach for multilingual SER using optimized multidomain acoustic representations. The research contributes toward improving SER systems for low-resource Indian languages and highlights the importance of feature diversity and optimization in multilingual emotional speech analysis.

Future work may focus on incorporating deep learning architectures, hybrid metaheuristic optimization techniques, attention-based models, and cross-corpus multilingual evaluation to further enhance generalization capability and real-world applicability of SER systems.

REFERENCES

[1] Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM, 61*(5), 90–99. https://doi.org/10.1145/3196520.

[2] Sultana, S., Iqbal, M. Z., Selim, M. R., Rashid, M. M., & Rahman, M. S. (2021). *Bangla speech emotion recognition and cross-lingual study using deep CNN and Bi LSTM networks*. IEEE Access, 10, 564–578.

[3] Chakraborty, C., Dash, T. K., Panda, G., & Solanki, S. S. (2022). Phase-based cepstral features for automatic speech emotion recognition of low resource Indian languages. *ACM Transactions on Asian Low-Resource Language Information Processing*, 1–16. https://doi.org/10.1145/3520415

[4] Koolagudi, S. G., Devliyal, S., Chawla, B., Barthwal, A., & Rao, K. S. (2012). Recognition of emotions from speech using excitation source features. *Procedia Engineering, 38*, 3409–3417. https://doi.org/10.1016/j.proeng.2012.06.39.

[5] Sultana, S., Rahman, M. S., Selim, M. R., & Iqbal, M. Z. (2021). SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *PLoS ONE, 16*(4), e0250173. https://doi.org/10.1371/journal.pone.0250173

[6] El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

[7] Koolagudi, S. G., Maity, S., Kumar, V. A., Chakrabarti, S., & Rao, K. S. (2009). IITKGP-SESC: Speech database for emotion analysis. In *IC3 2009, Communications in Computer and Information Science (Vol. 40, pp. 485–492).* Springer. https://doi.org/10.1007/978-3-642-03547-0_46

[8] Ali, S. A., Khan, A., & Bashir, N. (2015). Analyzing the impact of prosodic feature (pitch) on learning classifiers for speech emotion corpus. *International Journal of Information Technology and Computer Science, 2*, 54–59. https://doi.org/10.5815/ijitcs.2015.02.07

[9] Rao, K. S., & Koolagudi, S. G. (2011). Identification of Hindi dialects and emotions using spectral and prosodic features of speech. *Systemics, Cybernetics and Informatics, 9*(4).

[10] Swain, M., Maji, B., Kabisatpathy, P., & Routray, A. (2022). A DCRNN-based ensemble classifier for speech emotion recognition in Odia language. *Complex and Intelligent Systems, 8*, 4237–4239. https://doi.org/10.1007/s40747-022-00822-7.

[11] Joshi, D. D., & Zalte, M. B. (2013). Recognition of emotion from Marathi speech using MFCC and DWT algorithms. *International Journal of*

Advanced Research in Computer Engineering & Technology (IJARCET), 2(2), 2278–5140.

[12] Kamble, V. V., Deshmukh, R. R., Karwankar, A. R., Ratnaparkhe, V. R., & Annadate, S. A. (2015). Emotion recognition for instantaneous Marathi spoken words. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computation (FICTA 2014, Vol. 2, pp. 335–346). Advances in Intelligent Systems and Computing.*

[13] Rajisha, T. M., Sunija, A. P., & Riyas, K. S. (2015). Performance analysis of Malayalam language speech emotion recognition system using ANN/SVM. In *International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST)* (pp. 1097). Elsevier.

[14] Bansal, S., & Dev, A. (2015). Emotional Hindi speech: Feature extraction and classification. In *International Conference on Computing for Sustainable Development*.

[15] Nayak, B., Bisoyi, B., & Pattnaik, P. K. (2020). Smart emotion-based decision support system using machine learning technique. *International Journal of Advanced Research in Engineering and Technology, 11*, 2017–2079.

[16] Bharti, D., & Kukana, P. A hybrid machine learning model for emotion recognition from speech signals. In *IEEE International Conference on Smart Electronics and Communication (ICOSEC)*.

[17] Farhad, M., Ismail, H., Harous, S., Masud, M. M., & Beg, A. (2021). Analysis of emotion recognition from cross-lingual speech: Arabic, English, and Urdu. In *2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM 2021)* (pp. 42–47).

[18] Choudhury, N., & Sharma, U. (2021). Emotion recognition in standard spoken Assamese language using support vector machine and ensemble model. *Indian Journal of Computer Science and Engineering, 12*, 147–158.

[19] Mehra, P., & Verma, S. K. (2022). BERIS: An mBERT-based emotion recognition algorithm from Indian speech. *ACM Transactions on Asian Low-Resource Language Information Processing, 21*(5), Article 106, 1–19. https://doi.org/10.1145/3543956.

[20] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE Access, 9*, 47795–47814. https://doi.org/10.1109/ACCESS.2021.3068045

[21] Kandali, A. B., Routray, A., & Basu, T. K. (2008). Emotion recognition from Assamese speeches using MFCC features and GMM classifier. *TENCON 2008 – 2008 IEEE Region 10 Conference*, IEEE.

[22] Agrawal, S. S. (2011). Emotions in Hindi speech: Analysis, perception and recognition. In *IEEE International Conference on Speech Database and Assessments (Oriental COCOSDA)* (pp. 7–13). https://doi.org/10.1109/ICSDA.2011.6085972

[23] Agrawal, A., & Jain, A. (2020). Speech emotion recognition of Hindi speech using statistical and machine learning techniques. *Journal of Interdisciplinary Mathematics, 23*(1), 311–319. https://doi.org/10.1080/09720502.2020.1721926

[24] Pravena, D., & Govind, D. (2017). Development of simulated emotion speech database for excitation source analysis. *International Journal of Speech Technology, 20*, 327–338. https://doi.org/10.1007/s10772-017-9407-3

[25] Kannadaguli, P., & Bhat, V. (2018). A comparison of Bayesian and HMM based approaches in machine learning for emotion detection in native Kannada speaker. In *IEEE IEEMA Engineer Infinite Conference.*

[26] Dhar, A., & Shaikh, M. B. N. (2020). Emotion recognition with music using facial feature extraction and deep learning. In *Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST)* (pp. 1–4).

[27] Saad, F., Mahmud, H., Shaheen, M. A., Hasan, M. K., & Farastu, P. (2021). Is speech emotion recognition language independent? Analysis of English and Bangla languages using language-independent vocal features. *arXiv preprint arXiv:2111.10776*. https://arxiv.org/abs/2111.10776.

[28] Kumar, S., & Yadav, J. (2021). Emotion recognition in Hindi language using gender information, GMFCC, DMFCC and deep LSTM. In *ICMAI 2021, Journal of Physics: Conference Series* (pp. 1–11). https://doi.org/10.1088/1742-6596/1910/1/012015

[29] Fernandes, J. B., & Fernandes, J. B. (2022). Enhanced deep hierarchical GRU & BiLSTM using data augmentation and spatial features for Tamil emotional speech recognition. *I.J. Modern Education and Computer Science, 3*, 45–63.

[30] Emary, E., Zawbaa, H. M., & Hassanien, A. E. (2016) *"Binary grey wolf optimization approaches for feature selection"* Neurocomputing, Vol. 172, pp. 371–381. DOI: 10.1016/j.neucom.2015.06.083

[31] S. Mirjalili, S. M. Mirjalili, and A. Lewis, “Grey Wolf Optimizer,” Advances in Engineering Software, vol. 69, pp. 46–61, 2014.

[32] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[33] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536. https://doi.org/10.1038/323533a0 .

[34] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.