I.J. Image, Graphics and Signal Processing, **2025, 6, 109-132**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2025.06.07

Stacking Based Ensemble Learning with Deer Hunting Optimization for Automatic Identification of Malvani Dialects

Madhavi S. Pednekar*
Department of EXTC, Don Bosco Institute of Technology, Mumbai - 400070, Maharashtra, India
Email: madhavi@dbit.in
ORCID iD: https://orcid.org/0009-0001-1787-712X
*Corresponding Author

Kaustubh Bhattacharyya
Department of ECE, School of Technology, Assam Don Bosco University, Guwahati, Assam - 781017, India
Email: kaustubh.d.electronics@gmail.com
ORCID iD: https://orcid.org/0000-0002-9207-5020

Received: 11 January, 2025; Revised: 11 May, 2025; Accepted: 16 August, 2025; Published: 08 December, 2025

Abstract: Language Identification (LID) is a subset of Dialect Identification that addresses specific challenges and matters related to linguistic similarity between dialects. Various current approaches are used for dialect identification, but automated prediction is difficult because the clarity of voices is not in a perfect range, and inaccurate selection of features. It is essential to utilize an appropriate feature subset that contains sufficient signal information for the learning model to correctly recognize language dialects. So as to eradicate the mentioned issues, optimized stacking based ensemble learning is developed. The identification process initiates with the pre-processing by using an adaptive least mean square filter and a fractional bandpass filter. The features from the pre-processed audio signal will be extracted by using Gammatone frequency Cepstral coefficients (GFCC) and Shifted Delta Cepstral Coefficient (SDCC). Then, the extracted features will be reduced with the help of Independent Component Analysis (ICA). Furtherly, the classification of selected features will be further given to the Recurrent Neural Network (RNN), which acts as a meta-classifier and additionally gets information from a pair of distinct classifiers, such as Radial Basis Functional Neural Network (RBFNN) and Deep Belief Network (DBN). The hyperparameter present in the RNN classifier was tuned using the Deer Hunting Optimization Algorithm (DHOA). The proposed approach has an accuracy of 97%, a precision of 96%, also an F1-score of 97%. Therefore, for an automatic dialect identification, the suggested approach is the best option.

Index Terms: Dialect Identification, Audio Signal, Fractional Bandpass Filter, Gammatone Frequency Cepstral Coefficients, Shifted Delta Cepstral Coefficient, Deer Hunting Optimization

1. Introduction

Automatic language identification of speech is an automated technique for determining the language used in a digitized voice utterance. There are numerous methods for mechanically extracting data from a speech signal, including this one [1]. A single instance of a set of data types on how language identification can be used is pre-processing for human listeners [2]. A speech includes information on the speaker's content, pitch, voice speed, and other characteristics. It becomes evident that human interpersonal connections are significantly influenced by emotion. The Indo-Aryan language of Marathi is spoken by the Marathi community in Maharashtra. This is the authorized language of states in Maharashtra and Goa in Western India, one of the country's twenty-three official languages. Marathi has the fourth-highest native speaker population in India [3]. The language Marathi is one of the oldest Indo-Aryan languages still in use today, with literature dating back to about 900 AD. Standard Marathi and Varhadi are the two main dialects of Marathi. Among other languages, there are links between Khandeshi, Dangi, Vadvali, and Samavedi [4]. The advancement of human-machine interfaces will come from prying into users' emotions to ascertain their situations [5].

Even though some of the most widely spoken languages in the world already have systems in place for speaker detection, text-to-speech synthesis, automatic voice recognition, language identification, and other related functions, there are still no feasible speech technology alternatives for Indian languages [6]. The identification accuracies are considerably lower on out-of-domain data. Presents challenges for non-native speakers on the other side of the user end, which seems a bit difficult to gather or understand the dialect that the other user is using or speaking, also, there are a lot of threats to minority languages [7]. Speaking has several advantages that enable us to communicate information well beyond what we can express with words alone. Spoken language is defined as speaking aloud or using other comparable expressions with meaning to convey concepts or other information [8]. The main goal of language identification is to determine the language of the speech signal. In addition to these stated downsides, there is also the underlying model, which seems to become more complex [9]. Hence, to overcome these mentioned flaws, an automatic identification of Malvani dialects has been initiated.

Automatic Dialect identification has been proposed with a lot of existing methods, but accurate identification is difficult. Based on the Distinctive features in grammar, vocabulary, prosody, and pronunciation that can all be employed to differentiate across dialects, the existing methods have major drawbacks. The evacuation of noises frequently occurring, the propagation sector theme, loss of accurate prediction, bag-of-words model used correctly to predict lexical items, voice clarity is not perfect, and automated prediction models for language detection make it difficult to achieve an accurate level of result in the prediction rate because features are chosen incorrectly. It is also crucial to use a feature subset that is appropriate and contains enough signal information for the learning model to accurately identify dialects of languages. Oscillations in voice prediction and conversion seem hard and a bit delayed in identifying whether the language is Marathi or Malvani [10]. So, an optimized stacking based ensemble approach is developed. Various feature extraction approaches are used to extract spectral features, which are used to enhance the accuracy. The major contribution of this proposed work is listed below,

- Optimized stacking based ensemble learning is developed for identifying the Malvani dialects.
- Gammatone frequency Cepstral coefficients and shifted delta Cepstral coefficients are used to extract the spectral features.
- Features are selected by using Independent Component Analysis (ICA), which is used to improve the classification accuracy.
- RNN, RBFNN, and DBN are the three classifiers used in stacking based ensemble learning for classifying whether the language is Marathi or Marwari.
- The learning rate present in the RNN classifier was tuned using the Deer Hunting Optimization Algorithm.

The sections that follow are organized as follows: in the study, a broad range of research publications about the identification of contemporary dialects are included in Section 2. In Section 3, the description of the proposed methodology is clearly explained. The performance measures and results of the proposed framework are presented in Section 4. The research study is concluded in its entirety in Section 5.

2. Literature Review

Numerous recent studies are recognizing distinct languages of Malvani and Marathi dialect identification with different methods. A brief overview of a few of the recent methods is given.

Monterio et al. [11] had introduced an approach based on temporal pooling strategies with a deep neural network, named Insights into end-to-end learning scheme for language identification. This approach stated a dialect identification from the Arabic language. The major drawback was that the encoding of the signals that are used for identification purposes was not as much as in an expected effective resultant.

Shon et al. [12] had established an approach of unsupervised representation learning of speech for dialect identification for dialect identification (DID), a Factorized Hierarchical Variational Autoencoder (FHVAE) model combining the convolutional neural network was used to learn an unsupervised latent representation is implemented. The recognition of the language which has lots of dialects are as can be found out using this approached system. The downside of the approach is that the segmenting of the language seems to be a bit difficult.

Rajalakshmi et al. [13] had suggested a cross-lingual offensive language identification for low resource languages the cast of Marathi where automatic language identification in the Marathi and the malvani sectors. In this Indo-Aryan language identification of the Marathi. The drawback of the approach is that the accurate prediction of the malvani identification was not as much as satisfied in the identification.

Mustaqeem and Kwon [14] had established a CNN-Assisted Enhanced audio signal processing for speech emotion recognition was presented. Speech emotion categorization is done using a SoftMax classifier. To increase accuracy, the suggested method was tested using the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Interactive Emotional Dyadic Motion Capture (IEMOCAP) datasets. But, the accuracy attained here in this approach is not much in an efficient state of identification.

Sarma and Sarma [15] had created a neuro fuzzy classifier with feed forward neural network based classifier a system named dialect identification from speech using prosodic features and was implemented, which combines the

fuzzy classification and also resulted in an effective NFC based dialect identification within the languages. The drawback of this proposed approach is that the extreme low parameter usage for feature vectors the classifier's computing burden is quite little.

Ma et al. [16] had established a DNN based Automated Speech Recognition (ASR) for mixed dialect utterances by mixing dialects language models with autonomously determined weights that optimized the chance of recognition. The ASR based identification have been done in this proposed system is that the latter produced results with fewer errors by integrating recognition results with each dialect model for a single language.

Bhosale Rajkumar et al. [17] have developed a system with TIMIT and the tokenization process for the dialect identification name automated recognition of Latin American and Spanish conversational dialects. Here, the system has been trained to detect Spanish dialects spoken in Peru and Cuba, but it may be readily expanded to identify other dialects as well. The drawback is that the error comparison rate is in a bit high range.

Bharathi et al. [18] have suggested a Decision Tree (DT) for multilingual machine translation where the language identification. Using n-gram-based and machine learning-based language identifiers, three Indian languages such as Hindi, Marathi, and Sanskrit present in a document submitted for machine translation are identified following training. The drawback of this proposed work is that most of the instances are in noisy results.

Babhulgaonkar and Sonavane et al. [19] developed a machine learning n-gram based dialect identification from the audio based signals, which has been stated by automatically identifies a speaker's dialect from a dataset of distinct dialects based on a dialect transcript and transcript with audio recording. Only text is used in the first approach, while both text and audio are used in the second. The training, validation, and test splits were the same in both approaches. When it comes to dialect classification, a text-only model is less useful than one that combines text and audio is considered the major drawback.

Numerous approaches have been reviewed for automatic Malvani or Marathi identification has several downsides that are as the encoding of the signals that are used for identification purposes is not as much as in an expected effective resultant [11], segmenting of the language seems to be bit difficult [12],non-gain able accurate prediction, extreme low parameter usage for feature vector[13], errors by integrating recognition results with each dialect model for a single language [14], error comparison rate is in a bit high range [15], instances are in noisy result [16], noisy result [17], identification of dialect were ion not an appropriate rate [18], the converted audio was not in effective audio [19]. In order to overcome these issues the Stacking based Ensemble Learning is developed for automatic Identification of Malvani Dialects.

3. Proposed Methodology

Speaking is the act of actually putting thoughts or other information into words through meaningful speech or related utterances. However, it becomes very difficult to classify the languages spoken in multilingual countries like India because of idiomatic expressions, dialects, accents, and other linguistic quirks. However, because Indian languages are complex, using standard detection methods seems very difficult and inaccurate. Therefore, it is imperative that the language be automatically detected in order to determine whether or not the language is Malvani using stacking based ensemble learning. The automatic spoken language identification is shown in Figure 1.

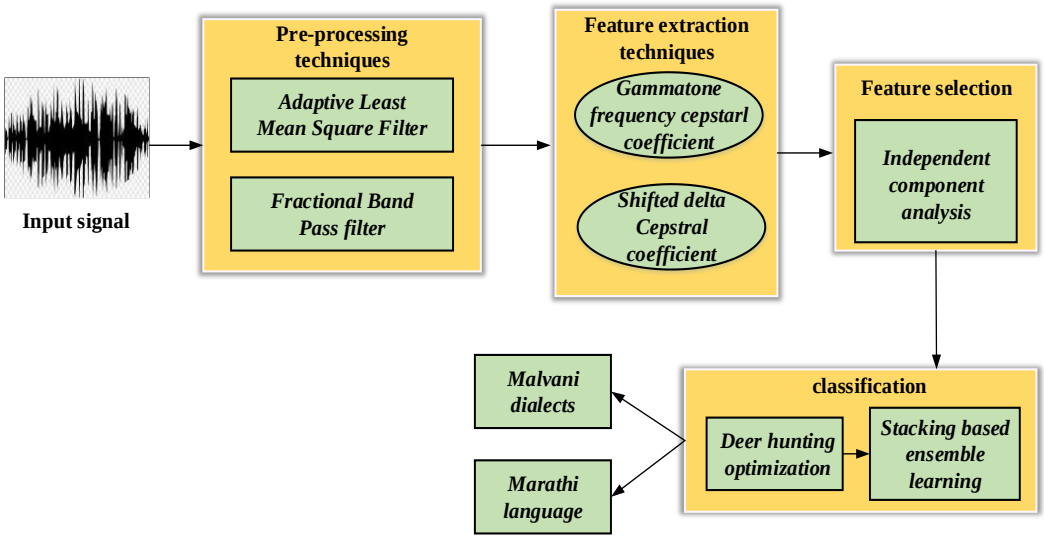


Fig. 1. Schematic architecture of proposed.

The proposed methodology states the Fractional band pass and adaptive least mean square filters are used for the pre-processing of raw audio signals. The output signal from the pre-processing step is given as an input to extract the

relevant feature. Gamma Frequency Cepstral Coefficients (GFCC) and Shifted Delta Cepstral Coefficient (SDCC) are accustomed to extract characteristics from the pre-processed signal. To eliminate unnecessary features and dimensions from the feature subset, Independent Component Analysis (ICA) was used to reduce the features' dimensions. Finally, a stacking based ensemble approach is developed to classify the language as malvani or Marathi. In this ensemble approach, a recurrent neural network (RNN) acts as a Meta classifier, which gets input from two other classifiers, such as a deep belief network (DBN) and a radial basis functional neural network (RBFNN). RNN classifier's learning rate was tuned through the application of the Deer Hunting Optimisation Algorithm (DHOA). The steps involved in the identification of the language, such as Marathi or Malvani are explained below.

Pre-Processing

Fractional band pass and adaptive least mean square filters are used for the pre-processing of raw audio signals. The original signal's noise is eliminated using an adaptive least mean square filter. For the purpose of dialect prediction, higher frequency bands are removed using a fractional bandpass filter. The pre-processing segment's specific method is outlined in more depth below.

Adaptive least mean square filter

Adaptive filters are typically employed in unpredictable time-variant environments or unfamiliar environments where preliminary identification can be difficult. This mean square filter employs a gradient descent technique where the instantaneous error signal is used to update the filter coefficients. Adaptive filters find many applications in noise reduction, equalization, and communication systems [20]. The purpose function of the adaptive least mean square filter is expressed in the Eqn. (1) below,

$$\xi(n \sum_{k=1}^{n} p_n(k) e_n^2(k) \tag{1}$$

$$p_n(k) = \lambda^{n-k} \tag{2}$$

Where parameter λ is indicated as a factor for forgetting in the range of 0 to 1, however it is advised to utilize values ranging from 0.95 to 1 and k = 1, 2, 3,... n. The output signal of the adaptive mean square filter is expressed as in the below Eqn (3).

$$y(n) = \sum_{i=0}^{N-1} w_i(n) x(n-i) \tag{3}$$

Fractional Bandpass filter

A particular frequency flows as an input signal that has been passed through a specific range of frequencies. In such cases, the fractional band frequency filters away signal frequencies that are excessively high or low in relation. At first, magnitude and phase charts are used to illustrate the active bandpass filter's response using integer-order elements which is given in the below Eqn. (4)

$$\frac{R_1}{R_1 + \frac{1}{C_1 S} + L_1 S} \tag{4}$$

Thus, the fractional band pass filter's transfer function is obtained throughout the above Eqn(5).

$$\frac{R_1}{R_1 + \frac{1}{C_1 j\omega} + L_1 j\omega} \{1 + \frac{R_f}{R_i}\} \tag{5}$$

Where the magnitude and the phase are compared with the transfer function which is shown in the above Eqn (6).

$$|mag| = \frac{R_1}{\sqrt{R_1^2 + (L_{1.\omega} - C_1^{-1}.\omega^{-1})^2}} \{1 + \frac{R_f}{R_i}\} \tag{6}$$

$$phase(\emptyset) = tan^{-1}\left(\left(L_{1.\omega} - C_1^{-1}.\omega^{-1}\right) \Big/ R_1\right) \tag{7}$$

The low frequency stop band attenuations are solely dependent on the order of the fractional capacitor, and the high frequency stop band attenuations are dependent only on the order of the fractional inductor. The stop band attenuations strongly correspond to the ones anticipated and demonstrate that the low and high frequency stop band attenuations are distinct from one another. Both scenarios are reproduced by displaying the magnitude and phase angles, where in case 1,

β is constant and α varies, and in case 2, α is constant and β is varied. The bandwidth exhibits nearly inverse variation with α, but the Q-factor gradually get decreased [21].

Feature Extraction

The output signal from the pre-processing step is given as an input to extract the relevant feature. The pre-processed input signal is divided into frequency sub-bands. In this feature extraction method, and then it is further normalized to produce feature coefficients. Gamma Frequency Cepstral Coefficients (GFCC) and Shifted Delta Cepstral Coefficient (SDCC) are used to extract characteristics from the pre-processed signal.

Gammatone Frequency Cepstral Coefficient

Noise is the primary issue with speech recognition. Additive noise signals are one of MFCC's main drawbacks. With the gammatone filter, one may simulate hearing. For non-speech audio categorization, these biologically inspired characteristics are applied using Gammatone filters with similar rectangular bandwidth bands. It is recommended to use the GFCC feature components to increase a speaker recognition system's dependability. One of the main drawbacks of MFCC is its sensitivity to additive noise. Based on a collection of Gammatone Filter banks, the Gammatone Frequency Cepstral Coefficients (GFCC) are an auditory characteristic. One can extract a Cochleagram, which is a frequency-time representation of the signal, from the Gammatone filter bank's output. They are derived from the commonly used Mel frequency cepstral coefficients (MFCCs). With a focus on low frequencies, GTCCs are specifically engineered to precisely represent and differentiate between non-speech audio signals. Noise is the primary issue with speech recognition. The additive noise signals of MFCC are one of its main drawbacks. The gammatone filter converts the nerve cells' impulse response into the auditory fibre, simulating the listening of spoken messages. Although every form represents the altered version of cell function, which is shown in the Eqn. (8)

$$g(t) = at^{n-1}e^{-2\pi b(f_c)t_{cos}}(2\pi f_c t + \emptyset) \tag{8}$$

Where t^{n-1} is the onset time, is the high (peak) value, e determines the bandwidth of the decay function, f_c sets the characteristic value and $\emptyset$ indicates the starting phase. GFCC coefficients, delta GFCC coefficients, and double delta GFCC coefficients make up the 36 total features produced for speech emotion recognition using the GFCC filter, which counts as 12.

The pre-emphasis of the second order filter lowers the dynamic range and emphasizes the frequency components that contain the majority of the essential information from the voice stream, as shown in the following Eqn(9).

$$H(z) = 1 + 4e^{-2\pi b/f_s}z^{-1} + e^{-2\pi b/f_s}z^{-2} \tag{9}$$

Similar to this a window covering K points is required for GFCC, which shifts each frame every L points. If fc is the centre frequency of the m-th filter, then each frame can be defined as $y(t; fc(m))$. Averaging $y(t; fc(m))$ across the window $t \in (nL, nL + K)$ yields the resulting Cochleagram representation for each frame, which has the following definition.

$$\bar{y}(n; m) = \frac{1}{k}\sum_0^{k-1} \gamma |y(nL + i; f_c(m))| \tag{10}$$

Where the magnitude of a complex number is represented by the other term and $\boldsymbol{\gamma}$ is a frequency dependent factor. Adding up $\bar{y}(n; m)$ for every channel results in yields.

$$\bar{y}\,(n) = |\bar{y}(n; 0), \ldots . \bar{y}(n; M-1)|^T \tag{11}$$

Where M is the filterbank's channel count for a 16 KH speech signal that results in 100 frames per second, typical recommended values are K = 400, L = 160, and M = 32. The matrix that emerges from this combination, $\bar{y}(n; m)$, is known as the Cochleagram.

Shifted delta Cepstral coefficient (SDCC)

The shifted delta coefficient (SDC) feature is typically used by statistical language recognition systems for automatic language recognition. SDCs are delta coefficients stacked across multiple frames. One popular feature extraction method for language recognition (LRE) is the Shifting Delta Cepstrum (SDC). A statistical language recognition system generally uses shifted delta coefficient (SDC) features for automatic language recognition. SDCs are stacked versions of the delta coefficient over several frames. Because it incorporates many frames, SDC performs better than typical delta and acceleration feature vectors and has a high context breadth. SDC is a transient characteristic that corporates the long-range dynamic features of the speech signal. The frame-based acoustic feature vectors are subjected to a shifting delta operation in order to produce new combined feature vectors for every speech frame that is provided. To determine the SDC coefficients at time t for a cepstral frame,

$$\Delta b_n(t,i) = b_n(t+iP+e) - b_n(t+iP-e); n = 0,1,\ldots,N_f-1, i = 0,\ldots,k_c-1 \quad (12)$$

Where P interval between subsequent delta computations is, i is the SDC block number, e is the lag of the deltas and cn is the nth cepstral coefficient. The final feature vector is produced by joining k_c blocks of N_f parameters together. The four variables that SDC coefficients are based on are typically represented as $N-e-P-k$. MFCCs are computed for every frame of data using the N dimension, that is, $b0, b1, \ldots, b_{N-1}$.

$$\Delta c(t+iP) = c(t+iP+d) - c(t+iP-d) \quad (13)$$

N-cepstral coefficients are computed, each frame, then for each of these cepstral streams the final vector at time t is given by the concatenation of the $Ac(t+iP)$ for all $0 < i < k$ [22].

Feature Selection

To further eliminate dimensions and unnecessary features from the feature subset, the extracted features are passed via a dimensionality reduction method. The Independent Component Analysis (ICA) was utilized to reduce the feature's dimension in order to enhance classification accuracy and decrease process complexity.

Independent Component Analysis

An algorithm that is crucial to the statistical signal processing field is independent component analysis. The detected signals are a linear combination of unidentified individual source signals since the technique is a linear transformation. Finding the inverted unknown mixing matrix is necessary before blindly estimating source signals. The following equation is the mathematical expression for the independent component analysis statistical model:

$$X = A.S \quad (14)$$

Where the mixed signals were included in the columns of the matrix $X \epsilon R^{M\times P}$, the number of rows indicates the various source signals, and the data in each column shows the number of samples of mixed signals. Where $S \epsilon R^{N\times P}$ matrix had source signal vectors in its columns, while the $A \epsilon R^{M\times N}$ matrix was a mixing matrix with elements that were unknown constants. This model demonstrates how combining S components yields the observed data (X). Since the elements of X are the only known data, A and S must be approximated using those data. The statistical independence of S components is the central ICA assumption. In mathematical terms, the joint probability function of all variables, if two or more are independent, is equal to the product of the probability densities of each variable. Data centring is the first step that must be taken while using ICA. to get every row in the X matrix to have a zero value, this entails reducing each element with the signal's mean value.

$$x_{ij} \leftarrow x_{ij} - \frac{\sum_{j'} x_{ij}}{P} \quad (15)$$

$$i = 1,\ldots,M, j = 1,\ldots,P \quad (16)$$

Additionally, S might be regarded as zero-mean. The cantered estimations are included with the mean vector of S following the estimation of the mixing matrix A. Whitening the observed variable is a further pre-processing procedure that eliminates any correlation in the data. Whitened is achieved by giving the X vector a linear transformation, which yields a new white vector $\widehat{X}$. The identity matrix is the covariance matrix of $\widehat{X}$:

$$E\{\widetilde{X}\ \tilde{X}^T\} = 1 \quad (17)$$

One well-liked whitening technique is used, which depends on eigenvalue breakdown on the covariance matrix of focused data.

$$E\{X\ X^T\} = E\ D\ E^T \quad (18)$$

Where $D$the diagonal matrix of eigenvalues and E is an orthogonal matrix of Eigenvectors. The most recent white vector $\widetilde{X}$ is going to be:

$$\widetilde{X} = ED^{-1/2}E^T X = ED^{-1/2}E^T\ AS = \tilde{A}S \quad (19)$$

The following iterative process is used to find a matrix W, also known as the un-mixing matrix, which is the inverse of the mixing matrix:

$$Y = WX \quad (20)$$

W is the unmixed matrix $N \times M$, where X is the mixture signals matrix $M \times P$ and Y is the extracted signals matrix $N \times P$. $W = A^{-1}$ and matrix A is invertible when the number of sources and mixed signals are equal. The problem is referred to as over-complete when A is not invertible and the number of mixes ($n > m$) is larger than the number of source signals [23]. The issue is referred to as under-complete and is resolved by eliminating some combination when the number of source signals is smaller than the number of mixtures ($n < m$).

Classification

The reduced features are given as input for the classification process. Here, Stacking based Ensemble Learning approach is employed for detecting whether the audio is malvani dialects or not. Recurrent Neural Network (RNN) acts as a Meta classifier and it receives data from two separate classifiers like Radial Basis Functional Neural Network (RBFNN) and Deep Belief Network (DBN).

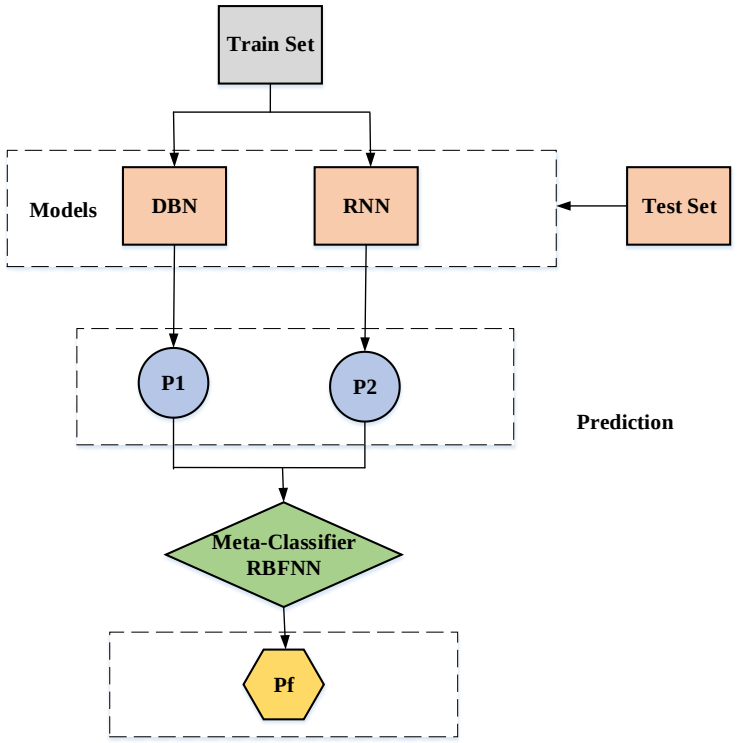


Fig. 2. Architecture of Stacking based Ensemble.

Figure 2 shows stacking based ensemble learning with the training sets, the test set, models, prediction, and meta-classifier. The stacking ensemble learning's architecture. The training set is used in the base-classifiers to train the layers and produce a fresh feature matrix for the meta-classifier layer. A suitable classifier is chosen by the meta-classifier layer to be classified for the final prediction.

After building the decision tree, the original GBDT algorithm reduces the tree by building a new one based on the empirical loss function's negative gradient. As opposed to this, the XGBoost includes regular terms when building the decision tree, which is described as

$$X_t = \sum_i l\left(y_i, F_{t-1}(x_i) + f_t(Y_i)\right) + \Omega(f_t) \tag{21}$$

Where $F_{t-1}(x_i)$ symbolizes the best course of action for $t-1$ trees. The definition of the normal word for the tree structure is

$$\Omega(f_t) = \gamma T + \tfrac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{22}$$

Where w_j the expected value of the ith leaf node and T is the total number of leaf nodes. It is evident that the XGBoost explicitly includes regular terms to regulate the model's complexity, which aids in preventing overfitting and enhances the model's capacity for generalization. The selected features will be ensemble by the radial bias functional neural network and the deep belief network [24].

Radial Bias Functional Neural Network

Radial Basis Functional Neural Network (RBFNN) is a two-layer network with a fully connected input layer to a hidden layer. The output of the hidden layer then calculates the output by adding the weighted sum of the input features. RBFNN uses a large number of Gaussian curves to learn how to approximate the underlying trend. The architecture of the RBFNN is shown in the Figure 3 below.

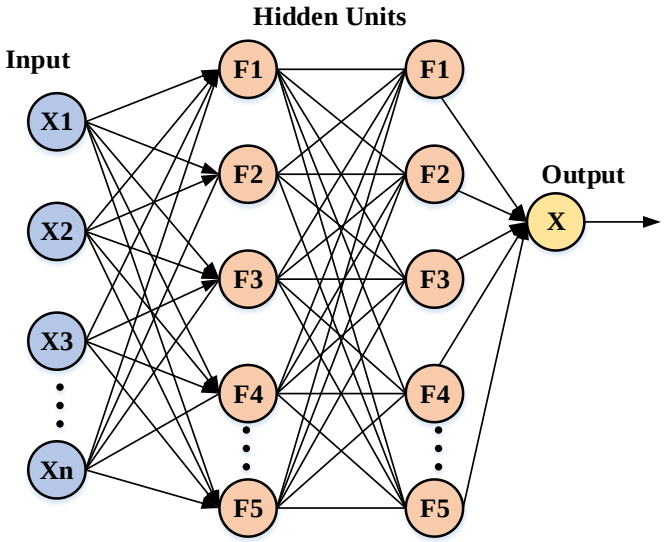


Fig. 3. Architectural diagram of Radial Bias Functional Neural Network.

The usage of Gaussian RBF inside the hidden layer neurons distinguishes RBF nets from conventional neural networks. RBF networks are employed in function approximation and regression. An RBF neural network is used to generate a weighted sum output given an input vector x.

$$F(x) = \sum_{j=1}^{k} w_j \varphi_j(x, c_j) + b \tag{23}$$

Where w_j is denoted as weights, b is the bias value, k is the number of centres and φ_j is considered as the Gaussian RBF, which can be calculated by the below given Eqn. (24).

$$\varphi_j(x, c_j) = e^{\left(\frac{-||x-c_j||^2}{2\sigma_j^2}\right)} \tag{24}$$

Where σ is known as the standard deviation, the basic actions for putting RBFNN into practice are Scale the features that were extracted. The vector mapping from input to a higher dimension. Determination of the final weights of the hidden layer. The final step is running a network test [25].

Deep Belief Network

A stack of RBM layers is appended to create a deep-belief network (DBN). Every RBM layer in the stack has communication capabilities with all levels, both preceding and following. As a result, it is a network that is put together using numerous single-layer networks. Every layer in a DBN, with the exception of the first and last layers, plays two roles: it acts as both the input layer for nodes that come after it and the concealed layer for nodes that come before it. The architecture of the Deep Belief Network is shown in Figure 4 below.

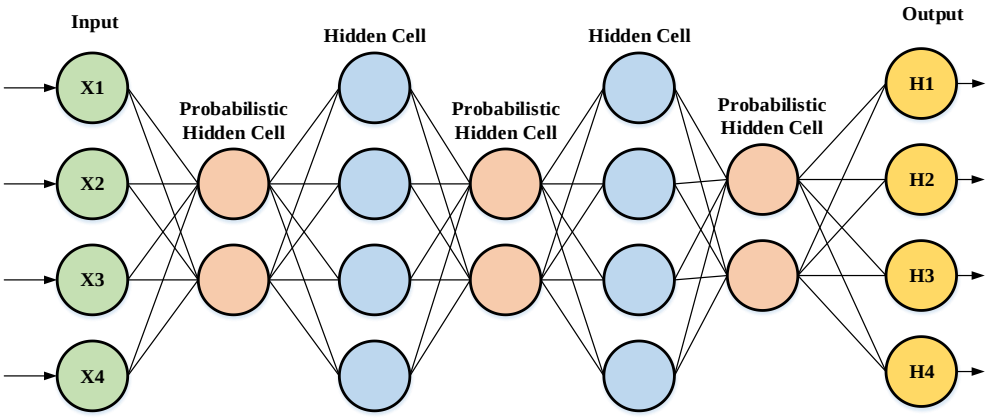


Fig. 4. Architectural Diagram of Deep Belief Network.

Deep belief networks can be used to recognize, group, and produce photos, video clips, and motion capture data, among other things. The combined distribution of the hidden layers HK and the observable vector x in the DBN model is shown in the Eqn. (25) as follows:

$$Pxh^1 \dots h^1 = \prod_{k=0}^{l-2} Ph^k h^{k+1} Ph^{l-1} h^1 \tag{25}$$

Where $Ph^{l-1}h^1$ is the visible-hidden joint distribution in the top-level RBM, $x = ho$, and $Ph^k h^{k+1}$ is a conditional distribution for the visible units conditioned on the hidden units of the RBM at level k. Furtherly, the classified features are correspondingly given to the recurrent neural network for the identification of the signal, whether that is a Marathi or Malvani dialects [26].

Recurrent Neural Network

Recurrent neural networks (RNNs) are better models to process sequential input and learn long temporal dependencies within the data. There are various RNN architectures, such as gated recurrent neural networks, long short term memory (LSTM) networks, and vanilla RNNs. The architecture of the recurrent neural network is shown in the below figure 5.

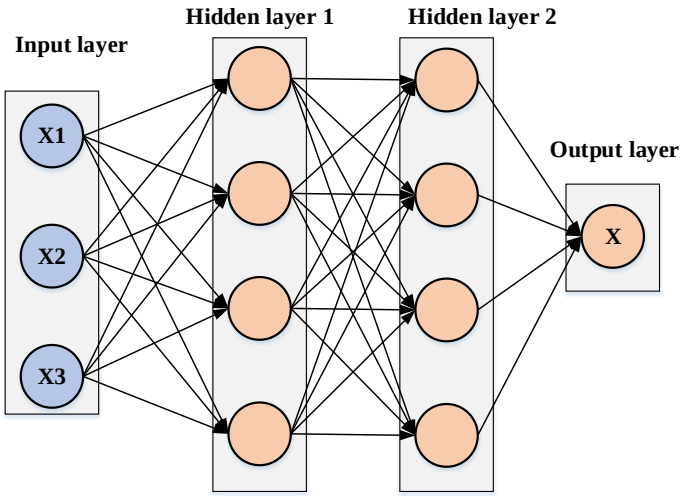


Fig. 5. Architectural Diagram of Recurrent Neural Network.

LSTMs and GRNs have the ability to simulate longer temporal relationships than vanilla RNNs. Since GRNNs are LSTM simplifications that achieve equivalent performance with fewer parameters.

$$h_t = \sigma^{(h)}(W^{(h)} \cdot h_{t-1} + W^{(x)} \cdot x_t \tag{26}$$

$$y_t = \sigma^{(y)}(W^{(y)} \cdot h_t) \tag{27}$$

$W^{(h)}$,$W^{(x)}$, and $W^{(y)}$ are the dense matrices that are optimized during training; on the other hand, $\sigma^{(h)}$ and $\sigma^{(y)}$ are nonlinear activation functions. When the procedure outlined is applied to every component of an input sequence,

$$\frac{\partial^2 u}{\partial t^2} - c^2 \cdot \nabla^2 u = f \tag{28}$$

$$h_t = A(h_{t-1}) \cdot h_{t-1} + P^{(i)} \cdot x_t \tag{29}$$

$$y_t = (P^{(0)} \cdot h_t)^2 \tag{30}$$

Thus, the gated recurrent unit (GRU) module, a variation of the RNN mod, receives the outputs from the batch normalization layers [27].

Deer Hunting Optimization Algorithm

The hunters engage in different activities; their approach to taking down the deer or buck depends on the hunting strategy they create. Owing to their unique talents, deer are able to flee from predators hunting. The movement of two hunters referred to as the leader and the successor in their optimal places determines the hunting strategy. The hunters surround the deer and approach it using a few different tactics in order to hunt it. The techniques take into account a number of factors, such as the direction of the wind, the location of the deer, and others. To that end, every hunter keeps his position updated until they get to the buck. Hunter cooperation is another essential component that makes hunting more effective. Ultimately, they arrive at their destination based on the roles of leader and follower. Effective management of deer requires a solid understanding of their behaviour. Additionally, deer behaviour can vary greatly depending on habitat, density of deer, and human disturbance. The characteristics of deer make hunting them more challenging. A crucial characteristic of deer is their vision. It has five times the strength of human vision in its visual features. The deer's ability to smell is another amazing trait. Deer sense is sixty times more sensitive than human sense of smell. A deer recognizes this as dangerous and lets out a loud snore while staggering around. Another deer can learn from is response. Deer are also adept at hearing the extremely high frequency of the noises.

Initialization

Initialize the learning rate based on the population of the hunters and the mathematical formulation is mentioned in the Eqn. (31)

$$Y = \{ y_1, y_2, \ldots, y_n \} \tag{31}$$

Fitness Function

DHO optimization approaches fitness function to minimize error values, which gives the mathematical function for defining the minimizing the error.

$$fitness\ function = minimization\ (E) \tag{32}$$

$$\mathrm{E} = \frac{1}{2} \sum_{\mathrm{m}=1}^{\mathrm{N}} (\mathrm{U_n} - \mathrm{V_n})^2 \tag{33}$$

The equation (32) states the error minimization. The Eqn. (33) has the values of $\mathrm{U_n}$ which is known as the expected value and $\mathrm{V_n}$ is known as the actual value.

Updation

The updation takes placed based on the deer vision power, sensibility and hearing power towards the hunter in these instance where the $S^w < 1$ meet this stage then it is said to be in the exploration phase and if the value lies in between the range of $S^w > 1$ which is said to be in the exploitation phase.

Exploitation Phase

There are two presumptive parameters which are the leader position (zl), which represents the hunter's initial optimal location, and the successor position (zs), which represents the hunter's position after that. The entire population attempts to update their location using the initial repetition to get into the optimal position. The mathematical formulation of the "encircling behaviour" is shown in the below Eqn. (34):

$$Z_{j+1} = Z_L - k \times S^w \times |L \times Z_L - Z_j| \tag{34}$$

The current and future locations are denoted by $Z_j|$ and Z_{j+1}, the random integer based on wind speed in the scope [0, 2] is shown byS^w, and the coefficient vectors are represented by L and k. Where γ is the random component and I_{max} represents the repetition peak in the interval [−1, 1]. The random integer in the range of 0 to 1 is denoted by δ. Determining the hunter's position requires careful consideration of angles. The prey shouldn't be able to see the successful attack, thus. The deer angle formula can be shown as follows in the Eqn. (35):

$$U_j = \frac{1}{8} \times \pi \times \alpha \tag{35}$$

The parameter that is taken into account for revising the angle of position is u, which results from the discrepancy that exists between the angle of the wind and the angle at which the prey is observed:

$$C_j = \beta_j - u_j \tag{36}$$

The following formula can be used to determine the new location once the angle of location has been determined,

$$Z_{j+1} = Z_j - S^w \times |\cos(\varphi_{j+1}) \times Z_l - Z_i| \tag{37}$$

Exploration Phase

The best solution is updated and made available at the leader location. The following exploratory updating formula is given in the below Eqn. (38):

$$Z_{j+1} = Z_s - k \times S^w \times |L \times Z_L - Z_j| \tag{38}$$

Where Z_s describes the hunters' successor location at any given time. The best answer in each iteration of this algorithm updates the hunters' location. The optimal solution is achieved when $|L| \geq 1$. If $|L| < 1$ the one of the hunters will be gets selected. Through the creation of L switch, this technique can change the algorithm's mode in between the stages of exploration and exploitation [28].

Termination

Once the optimal learning rate is selected, the process gets terminated. The flowchart of the proposed approach is shown in the below figure 6.

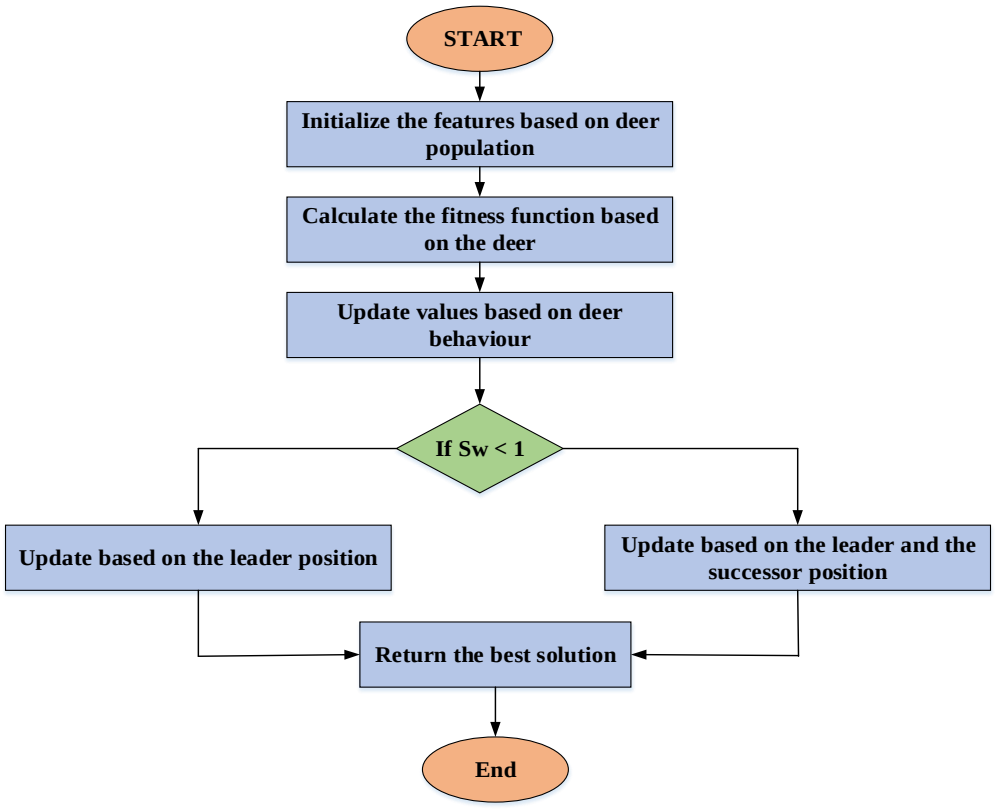


Fig. 6. Flowchart of Proposed System.

4. Result and Discussion

The proposed approach aim is to identify the malvani dialects identification through the stages of pre-processing by means of an adaptive least mean square filter, fractional bandpass filter. Then following that the extraction of the features will be happening through the Gammatone Frequency Cepstral Coefficient (GFCC) and Shifted Delta Cepstarl Coefficient (SDCC) then, the feature selection take place by the independent component analysis. The classification of the selected features will be further taken in to the Radial Bias Functional Neural Network (RBFNN) And Deep Belief Network (DBN). The features will be then given to the Recurrent Neural Network (RNN). Through this the dialects will be identified, whether Marathi or Malvani. Python is used to help with the testing 3.8 with CPU: Intel core i5, GPU: NVidia GeForce GTX 1650, 16-bit operating system, RAM: 16GB. The similarities of the parameters of RBFNN, DBN and RNN are shown in the below table 1, 2 and 3.

Table 1. Parameters of Radial Bias Functional Neural Network.

Parameters	Range
Learning rate	0.2
Momentum Factor	0.1
Maximum Iterations	500
Activation Function	Gaussian activation function

Table 2. Parameters of Deep Belief Network (DBN).

Parameters	Range
Number of epochs	10
Batch Size	100
Momentum for RBM	0
Learning rate alpha of RBM	1
Activation function of Neural Network	Sigmoid

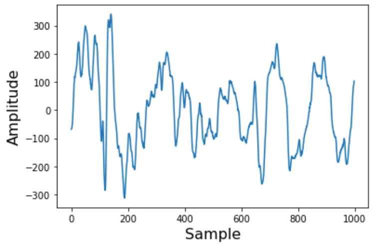
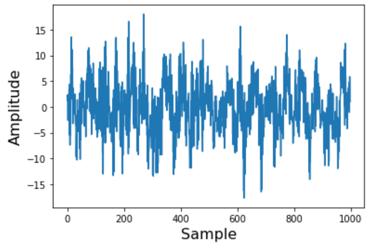
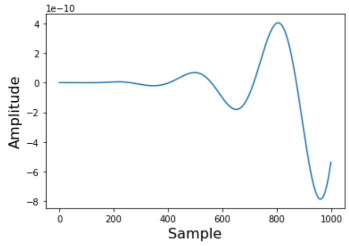
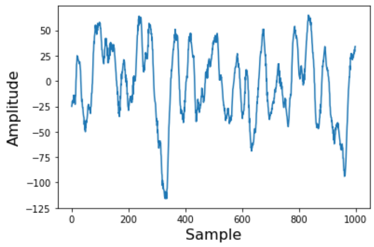
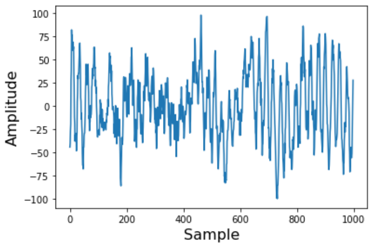
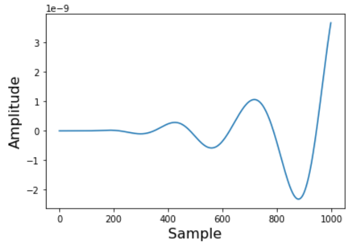
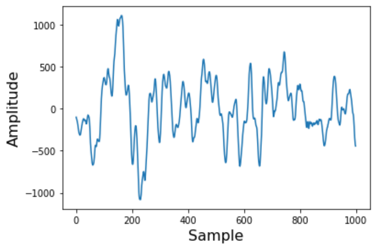
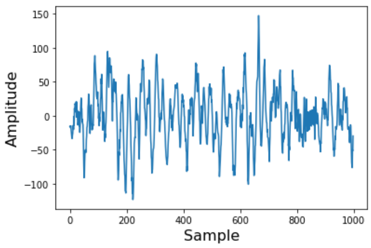
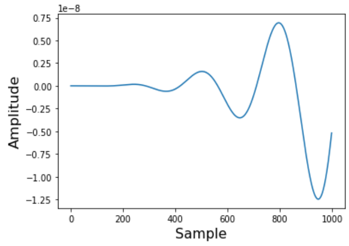
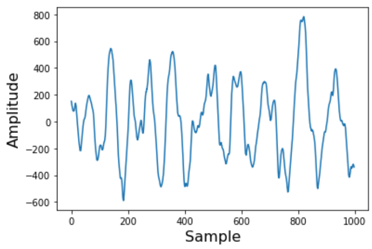
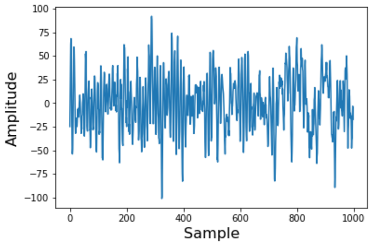
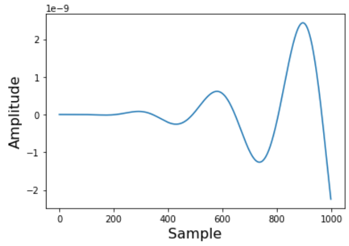
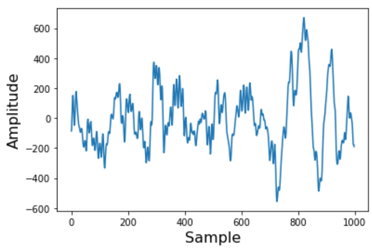
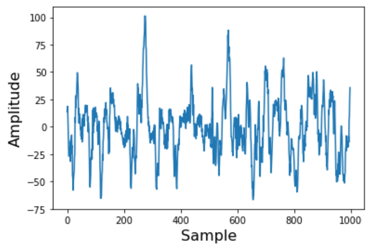
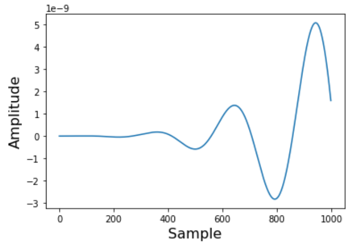
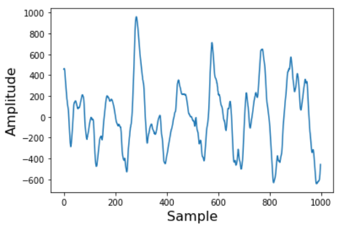
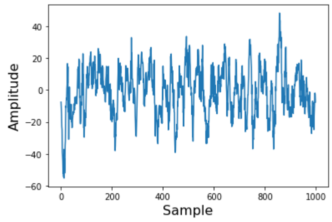
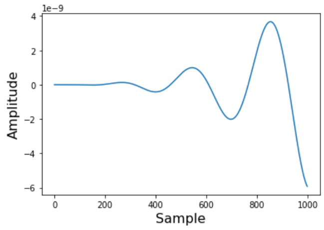
Table 3. Parameters for Recurrent Neural Network.

Parameters	Range
Learn rate	0.00002
momentum	0.8
Batch size	60
Halving factor	0.9

Dataset description

An automated identification system was developed for the Malvani dialect, a prominent variant of Marathi spoken in Maharashtra's Konkan region, supported by a comprehensive speech corpus totaling 16 hours and 25 minutes of spontaneous speech, including 4 hours and 10 minutes of clean recordings and 12 hours and 15 minutes of noisy data. Designed to capture the unique phonetic, prosodic, and linguistic features of Malvani, the corpus includes natural variations in tone, pitch, and context critical for accurate dialect recognition. The combined Malvani and standard Marathi speech database was divided into 70% training and 30% testing subsets to enable effective model learning and unbiased evaluation. This resource establishes a benchmark for Malvani dialect identification and contributes significantly to advancing dialect recognition and multilingual NLP research in India.

Initially, the existing data sets will be gathered and it will be going through the pre-processing phase using the adaptive least mean square filter and fractional bandpass filter, where the data's will be filtered and the noises will be removed at this step of pre-processing. The raw input and the pre-processed parameter of the simulation is shown in Figure 7.

Raw Input Signal	**Sample Image Pre-Processed Signal**	
	Adaptive Least Mean Square Filter	*Fractional Band Pass Filter*
		
		
		
		
		
		

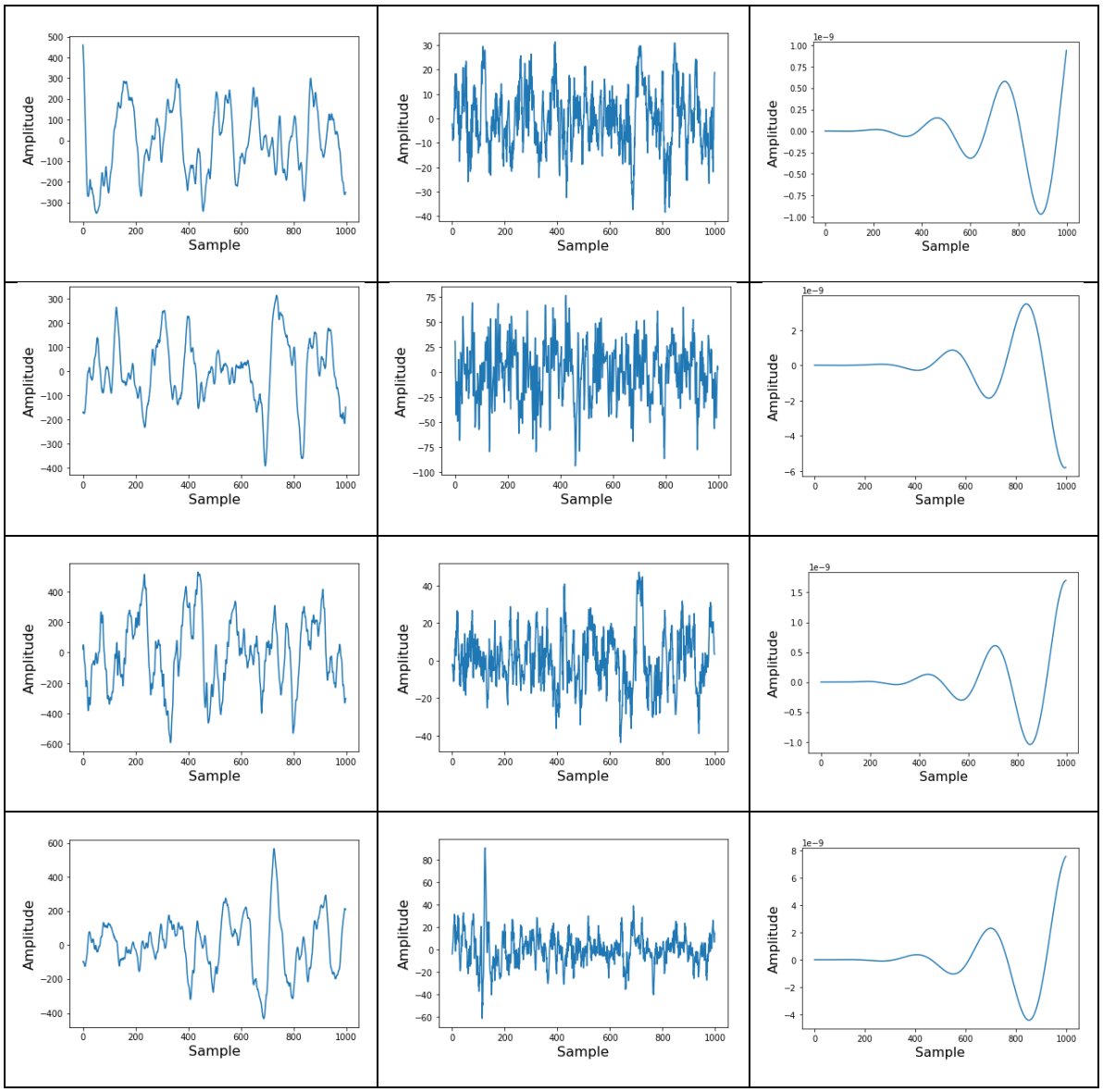


Fig. 7. Raw Signal and Sample Image Pre-Processed Signal.

Preferred methodologies such as Gamma Frequency Cepstral Coefficients (GFCC) and Shifted Delta Cepstral Coefficient (SDCC) for the extraction will be used during the ongoing process of feature selection and extraction. The audio signals will be categorized in the identification process in the last stage to determine whether the language is Marathi or Malvani.

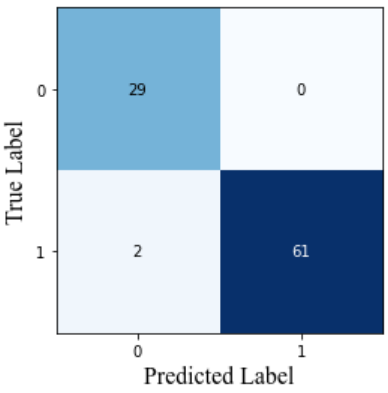


Fig. 8. Confusion matrix.

Figure 8 depicts the confusion matrix plot for the proposed system. The confusion matrix is used for evaluating the performance of a classification model which compares the actual target values with the predicted by the machine learning model. According to the above mentioned graph, the projected values of 29 samples for malvani (class 0) and 61 samples for Marathi (class 1).

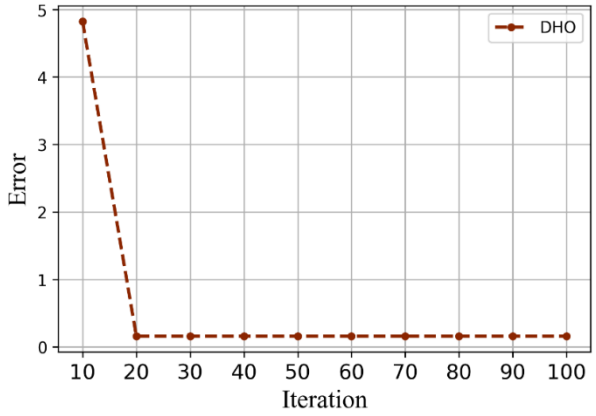


Fig. 9. ROC plot.

The convergence plot of the suggested and implemented optimizations is shown in Figure 9. Iteration with regard to the fitness function is typically used to generate the convergence plot. The fitness function is provided in this case as error. In this proposed optimization, the error value is 4.8 in the 20th iteration then it remains constant up to the 100th iteration.

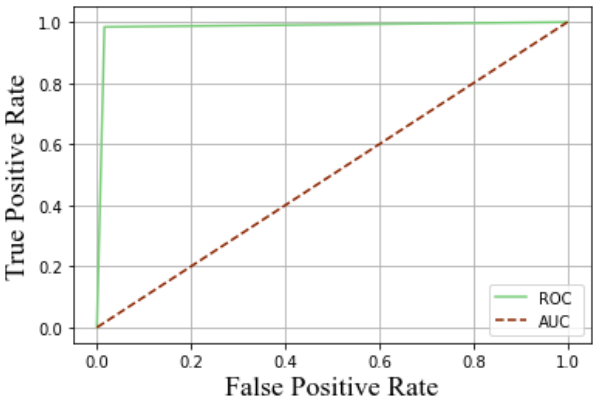


Fig. 10. Convergence Plot.

The methods of the proposed AUC and the ROC plots, which display the performance of the classification model at all classification levels, are shown in the aforementioned Figure 10. The proposed technique has a high degree of separability because the proposed approach's AUC is so unusually near to 1. The ROC curve shows how well the model can distinguish between classes.

Comparison Analysis

The comparison analysis beholds the existing techniques comparison where the methods involve several techniques named as accuracy, precision, recall, specificity, negative predictive value, false discover rate, f1- score, false negative ratio, false positive ratio, false omission rate, positive likelihood ratio, negative likelihood ratio, error, training time, testing time, execution time.

Figure 11 represents the graphical representation of the accuracy comparison with the existing approaches which has the following techniques named OSEL as 0.97%, DNN has the value of 0.94%, MLP has the range of 0.91%, ANN has the range of 0.75% and the RBF has the value of 0.85%. The precision comparison is shown in Figure 12 as a graphical representation which has the range of 0.96% for the OSEL, 0.93% for DNN, 0.98% for the MLP, 0.89% for the ANN and the RBF has the value of 0.84%.

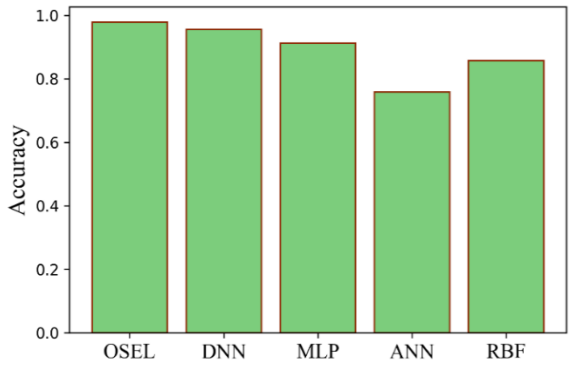


Fig. 11. Accuracy Comparison.

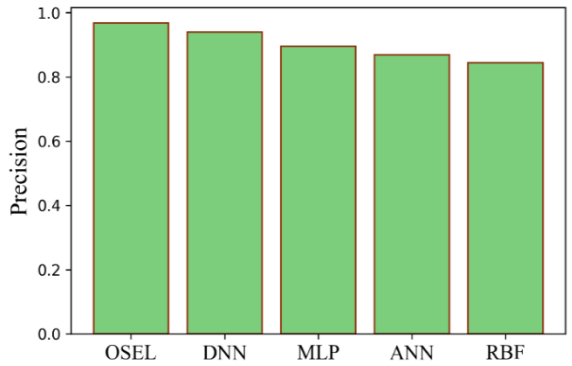


Fig. 12. Precision Comparison.

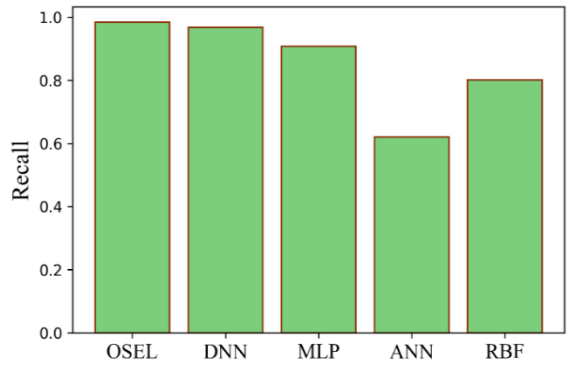


Fig. 13. Recall Comparison.

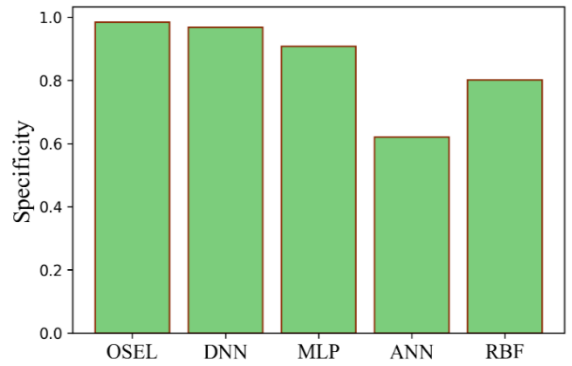


Fig. 14. Specificity Comparison.

This graphical depiction of the recall comparison with the existing techniques is shown in Figure 13. The following technical parameters are the OSEL has a value of 0.98%, the DNN has a value of 0.96%, the MLP has a range of 0.90%, the ANN has a range of 0.62%, and the RBF has a value of 0.80%. The picture displays a graphical representation of the specificity comparison, with values ranging from 0.98% for the OSEL, 0.96% for the DNN, 0.90% for the MLP, 0.89% for the ANN and 0.80% for the RBF shown in Figure 14.

Fig. 15. Negative Predictive Value Comparison.

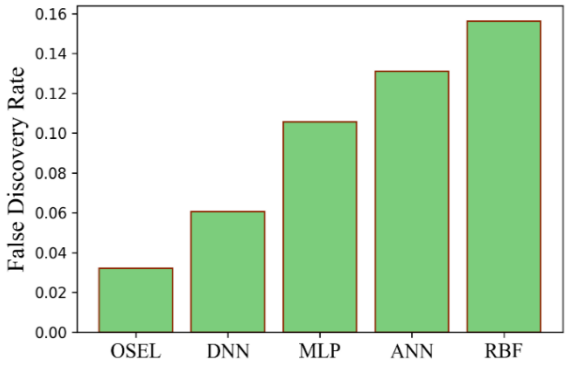


Fig. 16. False Discovery Rate.

The graphically representation of the Negative Predictive Value comparison is shown in Figure 15. OSEL has an negative predictive value comparison with existing approaches of 0.96%; DNN has a value of 0.93%; MLP has a range of 0.89%; ANN has a range of 0.86%; and RBF has a value of 0.75%. The false discover rate comparison is displayed graphically in Figure 16, with values ranging from 0.03% for the OSEL, 0.06% for the DNN, 0.10% for the MLP, 0.13% for the ANN and 0.15% for the RBF.

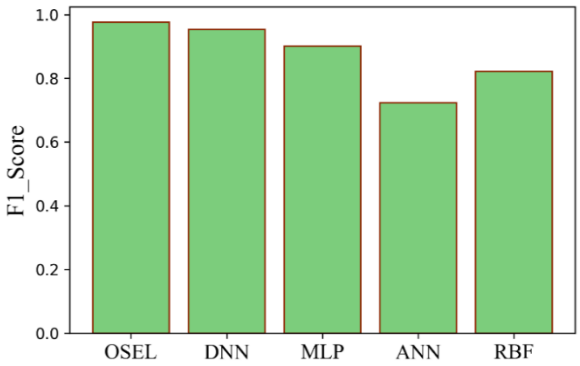


Fig. 17. F1-Score Comparison.

Figure 17 represents the graphical representation of the f1-score comparison with the existing approaches which has the following techniques named OSEL as 0.97%, DNN has the value of 0.95%, MLP has the range of 0.90%, ANN has the range of 0.72% and the RBF has the value of 0.82%. The false negative ratio comparison is shown in Figure 18 as a graphical representation which has the range of 0.01% for the OSEL, 0.03% for DNN, 0.09% for the MLP, 0.37% for the ANN and the RBF has the value of 0.19%.

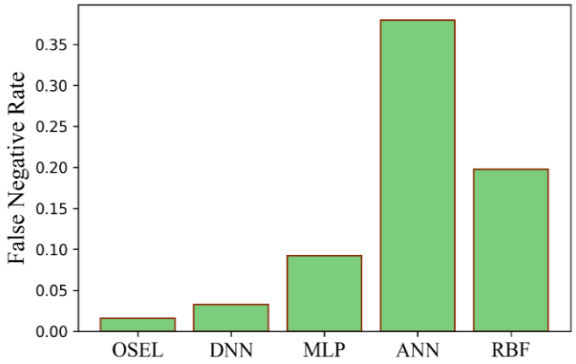


Fig. 18. False Negative Rate.

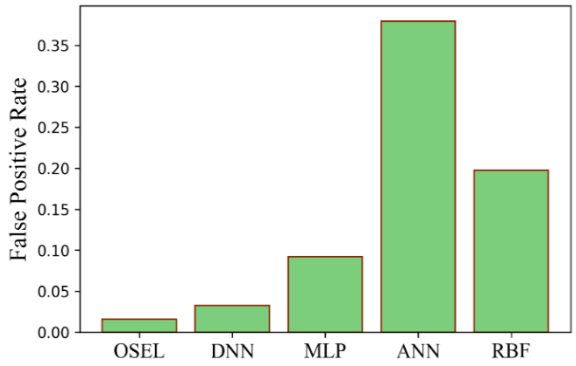


Fig. 19. False Positive Rate.

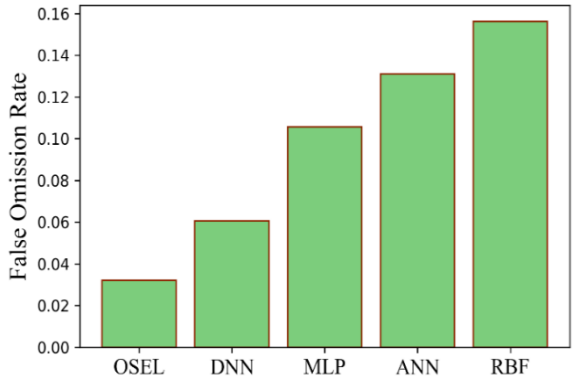


Fig. 20. False Omission Rate.

This graphical depiction of the false positive ratio comparison with the existing techniques is shown in Figure 19. The following technical parameters are the OSEL has a value of 0.01%, the DNN has a value of 0.03%, the MLP has a range of 0.09%, the ANN has a range of 0.37%, and the RBF has a value of 0.19%. The picture displays a graphical representation of the false omission rate comparison shown in figure 20, with values ranging from 0.03% for the OSEL, 0.06% for the DNN, 0.10% for the MLP, 0.13% for the ANN and 0.15% for the RBF.

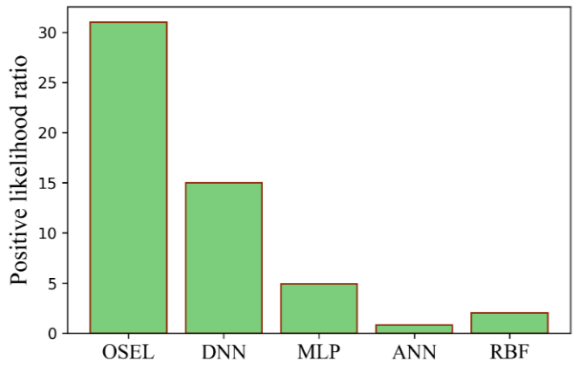


Fig. 21. Positive Likelihood Ratio.

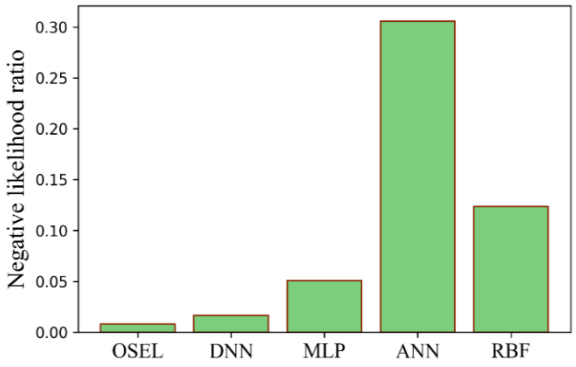


Fig. 22. Negative Likelihood Ratio.

The graphical representation of the Positive Likelihood Ratio is shown in Figure 21, where the OSEL has an positive likelihood ratio comparison with existing approaches of 31%; DNN has a value of 15%; MLP has a range of 4.93%; ANN has a range of 0.81%; and RBF has a value of 2.02%. The negative likelihood ratio comparison is displayed graphically in Figure 22, with values ranging from 0.008% for the OSEL, 0.016% for the DNN, 0.05% for the MLP, 0.30% for the ANN and 0.12% for the RBF.

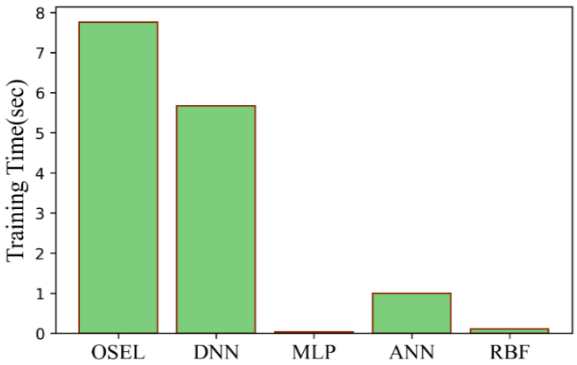


Fig. 23. Training Time.

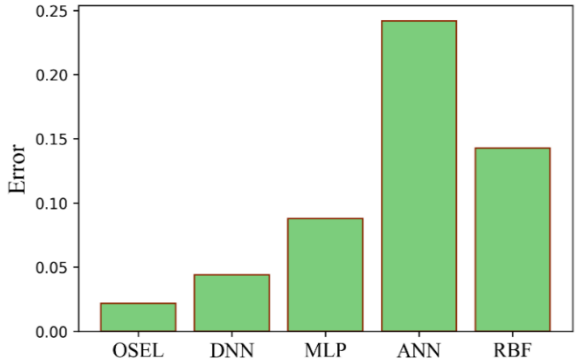


Fig. 24. Error Comparison.

This graphical depiction of the error comparison with the existing techniques is shown in Figure 23. The following technical parameters are the OSEL has a value of 0.24 sec, the DNN has a value of 0.14 sec, the MLP has a range of 0.08 sec, the ANN has a range of 0.37 sec, and the RBF has a value of 0.19 sec. Figure 24 displays a graphical representation of the training time comparison, with values ranging from 7.75% for the OSEL, 5.67% for the DNN, 0.03% for the MLP, 0.99% for the ANN and 0.10% for the RBF.

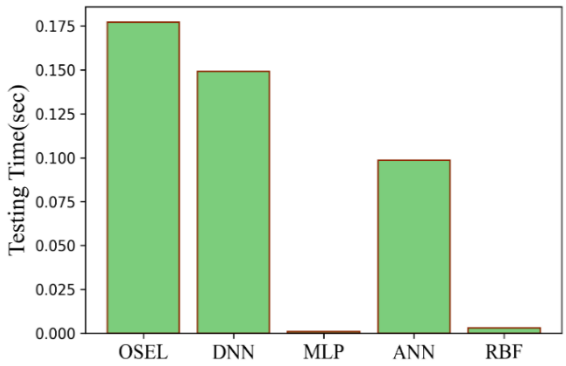


Fig. 25. Testing Time Comparison.

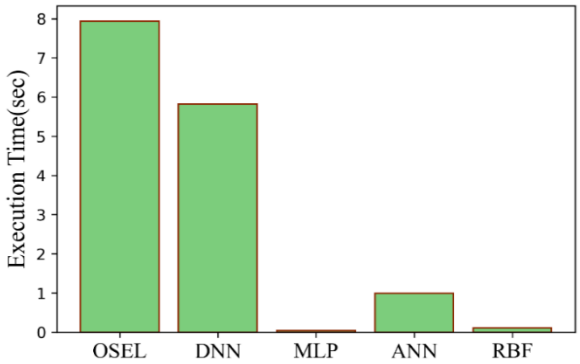


Fig. 26. Execution Time Comparison.

The following techniques are represented graphically in Figure 25, where the OSEL has a testing time comparison with existing approaches of 0.17 sec; DNN has a value of 0.14 sec; MLP has a range of 0.001 sec; ANN has a range of 0.09 sec; and RBF has a value of 0.003 sec. The execution time comparison is displayed graphically in Figure 26, with values ranging from 7.93 sec for the OSEL, 5.82 sec for the DNN, 0.03 sec for the MLP, 0.99 sec for the ANN and 0.10 sec for the RBF.

Error analysis

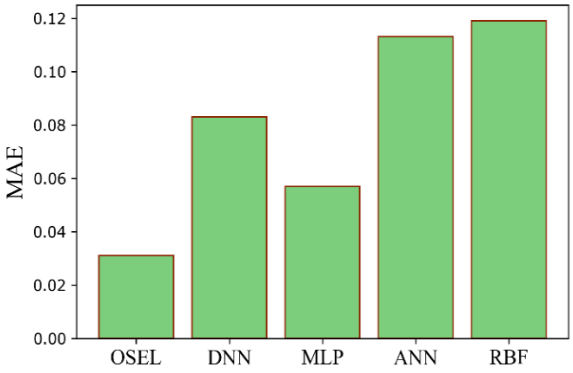


Fig. 27. Comparison of MAE

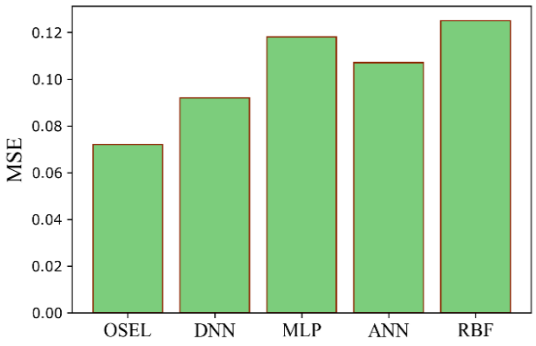


Fig. 28. Comparison of MSE

Figure 27 shows the comparison of mean absolute error (MAE). The proposed method achieve low MAE 0.031 with existing methods achieve DNN is 0.083, MLP is 0.057, ANN is 0.113 and RBF is 0.119. Mean Squared Error (MSE) comparison of proposed and existing method shows in figure 28. The proposed OSEL achieve low MSE value 0.072, other existing methods achieve DNN has 0.092, MLP has 0.118, ANN has 0.107, and RBF has 0.125. The proposed method achieve better values compared to other methods.

Table 4. Cross validation of the performance matrix

Performance	**70/30**	**60/40**	**50/50**	**80/20**
Accuracy	97%	95.6%	92.8%	96.3%
Precision	96%	94.8%	92.0%	95.4%
Recall	98%	95.0%	91.3%	96.1%
Specificity	98%	94.2%	91.0%	95.8%
NPV	96%	94.5%	91.7%	94.6%

The table 4 shows the model performance varies based on different training and testing data split ratios. A 70/30 split yields the highest scores across most metrics, suggesting better generalization and balanced training. This analysis helps determine the optimal train test spilt for balancing model training and validation.

Table 5. Statistical analysis for proposed method

	2 feature	**4 feature**	**6 feature**	**8 feature**	**10 feature**
F-value	621.929	622.061	622.082	621.875	621.932
P-value	0.058361	0.0087301	0.001954295	0.00017026	3.28562

This table 5 presents the statistical analysis results for the proposed method when using different numbers of features. F-values remaining consistently high, indicating stable model performance. The p-values decrease significantly as the number of features increases demonstrating improved statistical significance with more features.

Table 6. Comparative analysis of proposed with other optimization algorithms

Performance	**Proposed (DHOA)**	**PSO [29]**	**GA [30]**	**BO [31]**
Accuracy	97%	96.95%	68%	95.55%
F1- score	97%	-	67%	84.34%
Precision	96%	-	-	84.46%
Recall	98%	-	-	84.41%

Table 6 illustrate the comparison of proposed and other optimization algorithm. The existing techniques include Particle Swarm Optimization (PSO), Genetic Algorithms (GA), and Bayesian Optimization (BO). The proposed optimization achieve better performance compared to other optimization.

Table 7. Performance analysis of proposed and state of the art methods

Performance	**Proposed (OSEL)**	**DNN [32]**	**MLP [33]**	**ANN [34]**	**RBF [35]**
Accuracy	97%	96%	96.68%	87.5%	89%
Precision	96%	96%	97.27%	88.9%	-
Sensitivity	98%	-	96.79%	87.5%	-
F1-score	97%	-	-	87.5%	-

Performance analysis of proposed and state of the art methods comparison shown in table 7. The existing methods are Deep Neural Networks (DNN), Multilayer Perceptron (MLP), Artificial Neural Network (ANN) and Radial Basis Function (RBF). The proposed method achieve better performance compared to other methods.

5. Conclusion

Optimized Stacking Based Ensemble Learning (OSEL) approach is developed for identifying the malvani dialects. Dialect identification must be included in the audio logical test and clinical evaluation methods that are valid and reliable for native speakers of the target language which must be regularly used to examine speech perception abilities. Conventional speech recognition tests use most of the Marathi phonemes in their composition. The provision of low-frequency speech signals during a regular audio logical exam would present redundant information and overstate the performance of individuals with slanting high frequency hearing loss because of normal or nearly normal perception of these signals. There are lot of other approaches had been held and those approaches have lots of downsides such as extremely low parameter usage for feature vector, errors by integrating recognition results with each dialect model for a single language, error comparison rate a bit high range, instances are in noisy result, noisy result, identification of dialect were not at an appropriate rate, the converted audio was not in an effective audio stream, encoding of the signals that are used for the identification purposes is not as much as in an expected effective resultant. Thus, the adaptive least mean square filter (ALMSF) and the fractional band pass filter (FBF) will be used in the pre-processing phase to eliminate the Marathi dialect and the Malvani variants. The shifted delta Cepstral coefficient and the gamatone frequency Cepstral coefficients will next be used to extract the features. Constantly, the independent component analysis will have to be used to pick the extracted features. Afterwards, a stacking-based ensemble learning using a Deep Belief Network (DBN) and Radial Basis Functional Neural Network (RBFNN) combination continues. To classify the selected features, a recurrent neural network (RNN) functions as a meta classifier. The learning rate of the RNN classifier is adjusted using the Deer Hunting Optimization Algorithm (DHOA). The dialect identification system's average performance was determined to be an accuracy value has the range of 97%, precision value is of 96%, recall has the range of 98%, specificity value of 98%, Negative Predictive range is of 96%, False Discovery Rate is of 0.03%,f1-score is 97%, false negative ratio is of 0.015%, false positive ratio is 0.01%, false omission rate is of 0.03%, positive likelihood ratio is 31% negative likelihood ratio is 0.008 sec, error is of 0.02%, training time is of 7.75 sec, testing time is of 0.17 sec, execution time is of 7.93 sec. There have been attempts to find a solution to the challenge of distinguishing Malvani dialects in problem formulation. The suggested process yields significantly superior results than the existing methods, as the above graphical representation demonstrates.

Acknowledgement

The corresponding author claims the major contribution of the paper including formulation, analysis and editing. The co-author provides guidance to verify the analysis result and manuscript editing.

References

[1] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku. “GELP: GAN-excited linear prediction for speech synthesis from mel-spectrogram,” arXiv preprint arXiv:1904.03976. 2019.

[2] M. Hämäläinen, K. Alnajjar, N. Partanen, and J. Rueter. “Finnish dialect identification: The effect of audio and text,” arXiv preprint arXiv:2111.03800. 2021.

[3] R. Patil, B. Narkhede, S. Gaonkar, and T. Dave. “Deep Learning Based Marathi Sentence Recognition using Devnagari Character Identification,” In 2023 International Conference on Communication System, Computing and IT Applications (CSCITA). pp. 10-15, 2023. IEEE.

[4] M.S. Pramod. “A Critical Edition and English Translation of Manuscript Panchikarana (Doctoral dissertation, “Rajiv Gandhi University of Health Sciences (India)). 2020.

[5] S. Gaikwad, T. Ranasinghe, M. Zampieri, and C.M. Homan. “Cross-lingual offensive language identification for low resource languages: The case of Marathi,” arXiv preprint arXiv:2109.03552. 2021.

[6] K. Lounnas, H. Satori, M. Hamidi, H. Teffahi, M. Abbas, and M. Lichouri. “CLIASR: a combined automatic speech recognition and language identification system,” In 2020 1st international conference on innovative research in applied science, engineering and Technology (IRASET). pp. 1-5, 2020. IEEE.

[7] N. B. Chittaragi, and S. G. Koolagudi. “Dialect identification using chroma-spectral shape features with ensemble technique,” Computer Speech & Language. vol. 70, pp. 101230, 2021.

[8] S. Karthick, “Semi Supervised Hierarchy Forest Clustering and KNN Based Metric Learning Technique for Machine Learning System,” Journal of Advanced Research in Dynamical and Control Systems, vol. 9, pp. 2679-2690, 2017.

[9] S. Shivaprasad, and M. Sadanandam. “Identification of regional dialects of Telugu language using text independent speech processing models,” International Journal of Speech Technology. vol. 23, pp. 251-258, 2020.

[10] N. B. Abdallah, S. Kchaou, and F. Bougares. “Text and speech-based tunisian Arabic sub-dialects identification,” In Proceedings of The 12th Language Resources and Evaluation Conference. pp. 6405-6411, 2020.

[11] J. Monteiro, M. J. Alam, and T. Falk. “On the performance of time-pooling strategies for end-to-end spoken language identification,” In Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 3566-3572, 2020.

[12] S. Shon, A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu, and K. J. Han. “Slue: New benchmark tasks for spoken language understanding evaluation on natural speech,” In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7927-7931, 2022. IEEE.

[13] R. Rajalakshmi, F. Mattins, S. Srivarshan, and L.P. Reddy. "Hate Speech and Offensive Content Identification in Hindi and Marathi Language Tweets using Ensemble Techniques," In CEUR Workshop Proceedings. pp. 1-11, 2021.
[14] Mustaqeem, and S. Kwon. "A CNN-assisted enhanced audio signal processing for speech emotion recognition," Sensors. vol. 20, no. 1, pp. 183, 2019.
[15] M. Sarma, and K.K. Sarma. "Dialect identification from Assamese speech using prosodic features and a neuro fuzzy classifier," In 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN). pp. 127-132, 2016, IEEE.
[16] Z. Ma. "Speech processing with deep learning for voice-based respiratory diagnosis: a thesis presented in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science at Massey University, Albany, New Zealand (Doctoral dissertation, Massey University)," 2022.
[17] S. Bhosale Rajkumar. "A Holistic Review of Automatic Speech Recognition Systems for Real-time Implementation," Mathematical Statistician and Engineering Applications. vol. 71, no. 4, pp. 12341-12359, 2022.
[18] B. Bharathi. "SSNCSE_NLP@ DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text," In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages. pp. 313-318, 2021.
[19] A. Babhulgaonkar, and S. Sonavane. "Language identification for multilingual machine translation," In 2020 International Conference on Communication and Signal Processing (ICCSP). pp. 401-405, 2020. IEEE.
[20] R. Martinek, J. Rzidky, R. Jaros, P. Bilik, and M. Ladrova. "Least mean squares and recursive least squares algorithms for total harmonic distortion reduction using shunt active power filter control," Energies. vol. 12, no. 8, pp. 1545, 2019.
[21] K. Biswal, S. Swain, M. C. Tripathy, and S. K. Kar. Modeling and performance improvement of fractional-order band-pass filter using fractional elements. IETE Journal of Research. vol. 69, no. 5, pp. 2791-2800, 2023.
[22] J. Basu, S. Khan, R. Roy, T. K. Basu, and S. Majumder. "Multilingual speech corpus in low-resource eastern and northeastern Indian languages for speaker and language identification," Circuits, Systems, and Signal Processing. vol. 40, pp. 4986-5013, 2021.
[23] D. IRIMIA, and E. C. BOBRIC. APLICATION OF INDEPENDENT COMPONENT ANALYSIS IN LOAD PROFILE STUDY.
[24] T. T. Khoei, M. C. Labuhn, T. D. Caleb, W. C. Hu, and N. Kaabouch. "A stacking-based ensemble learning model with genetic algorithm for detecting early stages of Alzheimer's disease," In 2021 IEEE International Conference on Electro Information Technology (EIT). Pp. 215-222, 2021, IEEE.
[25] M. H. Bhatti, J. Khan, M. U. G. Khan, R. Iqbal, M. Aloqaily, Y. Jararweh, and B. Gupta, "Soft computing-based EEG classification by optimal feature selection and neural networks," IEEE Transactions on Industrial Informatics. vol. 15, no. 10, pp. 5747-5754, 2019.
[26] H. Zhao, J. Liu, H. Chen, J. Chen, Y. Li, J. Xu, and W. Deng, "Intelligent diagnosis using continuous wavelet transform and gauss convolutional deep belief network," IEEE Transactions on Reliability, 2022.
[27] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F. M. Smolle-Jüttner, H. Olschewski, and F. Pernkopf. "Multi-channel lung sound classification with convolutional recurrent neural networks," Computers in Biology and Medicine. vol. 122, pp. 103831, 2020.
[28] W. Ha, and Z. Vahedi. "Automatic breast tumor diagnosis in MRI based on a hybrid CNN and feature-based method using improved deer hunting optimization algorithm," Computational Intelligence and Neuroscience, 2021.
[29] Kukana, P., Sharma, P., & Bhardwaj, N. (2024). Optimized featured swarm convolutional neural network (OFSCNN) model based dialect recognition system for Bagri Rajasthani language. *International Journal of Information Technology*, 1-13.
[30] Ibrahim, A. B., Seddiq, Y. M., Meftah, A. H., Alghamdi, M., Selouani, S. A., Qamhan, M. A., ... & Alshebeili, S. A. (2020). Optimizing arabic speech distinctive phonetic features and phoneme recognition using genetic algorithm. *IEEE Access*, *8*, 200395-200411.
[31] Kumar, G., & Bhardwaj, S. (2025). Biomimetic Computing for Efficient Spoken Language Identification. *Biomimetics*, *10*(5), 316.
[32] Pednekar, M. S., & Bhattacharyya, K. (2024). Automatic identification of Malvani dialects from audio signal based on hybrid FFO-TSO with deep neural network. *Multimedia Tools and Applications*, 1-33.
[33] Barfungpa, S. P., Samantaray, L., Sarma, H. K. D., Panda, R., & Abraham, A. (2023). Dt-SNE: Predicting heart disease based on hyper parameter tuned MLP. *Biomedical Signal Processing and Control*, *86*, 105129.
[34] Wijonarko, P., & Zahra, A. (2022). Spoken language identification on 4 Indonesian local languages using deep learning. *Bulletin of Electrical Engineering and Informatics*, *11*(6), 3288-3293.
[35] Muttaqi, M., Degirmenci, A., & Karal, O. (2022, September). US accent recognition using machine learning methods. In *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-6). IEEE.

Authors' Profiles

Madhavi Pednekar holds a doctoral degree in Automatic Dialect Identification and brings with her over 23 years of dedicated experience in the field of education, all of which she has proudly served at Don Bosco Institute of Technology (DBIT), Mumbai. She currently serves as the Head of the Department of Electronics & Telecommunication Engineering at DBIT. With a strong academic and research background, she has published more than 20 research papers in reputed international journals. Her core areas of interest include Signal Processing, Speech Processing, and Embedded Systems.

In addition to her academic responsibilities, she also serves as the NSS Lady Program Officer, representing DBIT under the University of Mumbai, where she actively fosters a spirit of community service and social responsibility among students.

Kaustubh Bhattacharyya is currently working as an Associate Professor and Head in the Department of Electronics & Communication Engineering at Assam Don Bosco University, Assam, India. He earned his PhD. in Electronics and Communication Engineering from Assam Don Bosco University, and MTech in Electronics and Communication Technology, MPhil in Electronics Science, MSc. in Electronics Science and BSc in Electronics from Guwahati University, Assam. His research area includes High Frequency Communication System, VLSI, Nanotechnology, AI and ML. He holds vast research experience as a PhD guide and by work in various funded projects. He has published many book chapters, journal papers and conference papers.

How to cite this paper: Madhavi S. Pednekar, Kaustubh Bhattacharyya, "Stacking Based Ensemble Learning with Deer Hunting Optimization for Automatic Identification of Malvani Dialects", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.17, No.6, pp. 109-132, 2025. DOI:10.5815/ijigsp.2025.06.07