INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Improved Support Vector-Recurrent Neural Network with Optimal Feature Selection-based Spoken Language Identification System

Irshad Ahmad Thukroo[1], **Rumaan Bashir**[2]*, **J Kaiser Giri**[2]

1 Research Scholar, Department of Computer Science Islamic University of Science & Technology, Kashmir
2 Associate Professor, Department of Computer Science Islamic University of Science & Technology, Kashmir

OPEN ACCESS

Received: 02-11-2022
Accepted: 09-12-2022
Published: 07-03-2023

Citation: Thukroo IA, Bashir R, Giri JK (2023) Improved Support Vector-Recurrent Neural Network with Optimal Feature Selection-based Spoken Language Identification System. Indian Journal of Science and Technology 16(10): 680-697. https://doi.org/10.17485/IJST/v16i10.2119

* **Corresponding author**.

rumaan.bashir@islamicuniversity.edu.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objective: Spoken language identification being the fore-front of language recognition tasks and most significant medium of communication has to be enhanced in order to improve the accuracy of recently developed spoken language recognition systems. The purpose of this paper is to enhance the Spoken Language Identification (SLID) model using hybrid machine learning with deep learning model for regionally spoken languages of Jammu & Kashmir (JK) and Ladakh. **Method:** Initially, the speech signals of different languages of JK and Ladakh are manually collected from diverse sources, and it is pre-processed using Spectral Noise Gate (SNG) filtering technique. Once the speech signals are pre-processed, the feature extraction is performed by the cepstral features like Mel-frequency Cepstral Coefficients (MFCCs), Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP), and spectral features like spectral roll off, spectral flatness. **Findings:** From this feature extraction, the length of the feature vector seems to be long, and it is required to reduce the feature length. Hence, optimal feature selection is accomplished using the new meta-heuristic algorithm termed Adaptive Distance-based Tunicate Swarm Algorithm (AD-TSA) by considering the minimum correlation as objective. Finally, the language identification is handled by the hybrid classifier termed Improved Support Vector Machine-Recurrent Neural Network (ISVM-RNN). **Novelty:** The identification learning algorithm is enhanced by the AD-TSA by considering the minimum correlation as objective among features in order to get minimum number of features that are sufficient for language identification process. The efficiency of the proposed hybrid approach is validated by simulating the experiment on a user-defined language database of JK and Ladakh speech signals in the working platform of Python.

Keywords: Language Identification; Kashmir Languages; Optimal Feature Selection; Improved Support Vector MachineRecurrent Neural Network; Adaptive DistanceBased Tunicate Swarm Algorithm

1 Introduction

SLID is a process of recognizing the spoken languages from a speech utterance. In speech recognition, the language of the spoken segments can be determined by the language recognizers. The SLID on regional languages is investigated to widen the range of technology. However, the hybrid features-based language identification system performance has not been explored much. In 2021, Garain et al.[1] have developed a ensemble architecture known as Fuzzy GCP (Google Cloud Platform) based on deep learning to identify the spoken languages from speech signals. Several classification techniques such as Deep Convolutional Neural Network (DCNN), Semi-supervised Generative Adversarial Network (SSGAN), and Deep Dumb Multi-Layer Perceptron (DDMLP) were combined to attain maximum precision. In 2020, Deshwal et al.[2] have implemented a robust technique for language identification system based on hybrid features extraction. Various techniques were employed individually in the phase of feature extraction such as Perceptual Linear Prediction features (PLP), Mel Frequency Cepstral Coefficients (MFCCs), Relative Spectral Transform Perceptual Linear Prediction features (RASTA-PLP). In 2020, Albadr et al.[3] have proposed a new improved optimization technique namely Particle Swarm Optimization–Extreme Learning Machine (PSO–ELM)-based on Enhanced Self-Adjusting Extreme Learning Machine (ESA-ELM) technique. In 2019, Ma et al.[4] have presented a novel approach for SLID based on en-to-end short utterances, which was only suitable to identify short utterances. The LSTM method with transfer learning was used in this approach to extract the features. In 2020, Das et al.[5] have developed a novel model based on nature-inspired feature selection algorithm by combining the hybrid methods of Late Acceptance Hill-Climbing (LAHC) and Binary Bat Algorithm (BBA). This recognition method has also included the extracted features from audio signals by deep learning architectures such as i-vector and x-vector as well as Gammatone Frequency Cepstral Coefficient (GFCC), MFCC, Linear Prediction Coefficient (LPC), and Discrete Wavelet Transform (DWT). The initial results were compared with the results produced by the combination of both LPC and MFCC. Among them, the combined results have provided better performance by reducing complexities of the existing models. The proposed model has achieved high accuracy by using the Random Forest (RF) classifier. In 2021, Aarti & Sunil[6] proposed a hybrid SLID system for nine Indian languages for the training and testing utterance duration mismatch conditions. Initially Random forest-based importance vectors of 1582 OpenSMILE features were calculated for each utterance in different duration datasets. The major drawback of this model is that it worked on recorded speech only as testing was not done on any other dataset and the number of working language were limited to only nine languages that doesn't include low resource languages like Kashmiri. Ladakhi etc. In 2022, Alashban et al.[7] proposed a spoken language identification system that depends on the sequence of feature vectors. The proposed system used a hybrid Convolutional Recurrent Neural Network (CRNN) that combines a Convolutional Neural Network (CNN) with a Recurrent Neural Network (RNN) network, for spoken language identification on seven languages, including Arabic, chosen from subsets of the Mozilla Common Voice (MCV) corpus. Since the proposed model worked on Non-Indian languages by focusing on hybrid approach to gain advantage of two classifiers with hybrid feature. Though it increased the accuracy but the experimental cost also increased. In 2021, Gundeep et al.[8] proposed a SLID system that used audio files and converts those files into spectrogram images. It applies the convolutional neural network (CNN) to bring out main attributes or features to detect output easily. The main objective was to detect languages out of English, French, Spanish, and German, Estonian, Tamil, Mandarin, Turkish, Chinese, Arabic, Hindi, Indonesian, Portuguese, Japanese, Latin, Dutch, Pushto, Romanian, Korean, Russian, Swedish, Tai and Urdu. Experiments were conducted on different audio files using the Kaggle dataset . By testing the model on speaker independent data the accuracy may drastically fall. In 2021, Thukroo and Bashir[9] proposed a Mel-spectrogram-based approach using convoluttiosn[9] s neural networks for regionally spoken languages of Jammu & Kashmir and Ladakhi. The dataset contains six languages i.e. Kashmiri, Ladakhi, Urdu, English, Hindi, and Dogri. Initially, speech segments were converted into Mel-spectrograms by using inverse Fourier transformation to log of Fourier transformation of a time-domain signal, and at the backend, CNN serves as a classifier. Experiments were conducted on recorded speech, IIIT-H, and VoxForge. It found that an average accuracy of 100 % was achieved by running the model at 100 epochs. One of the main drawbacks of this model is speech segments were converted into image domain, therefore, the focus has been shifted to image domain rather than linguistic characters such as syntax and semantics of the language. Second testing was done by using speaker-dependent samples instead of speaker-independent samples. Third, the effect of noise has not been tested in a real domain. From many deep learning-based spoken language recognition models, it is observed that the method fails when the Indic languages are identified, which have various similar properties among them, and the same model is employed in both foreign and Indic languages and may give non-optimal results. Many techniques were developed in recent years for the spoken language identification, as depicted in the Table **??** DDMLP, DCNN, and SSGAN[1] help to attain the precision to the maximum level. Although, if the separate classes other than the training data classes are used, then the performance will be drastically decreased because of the dependency on the parameter choice. MFCC, PLP and RASTA-PLP[2] provide high accuracy and minimum error in the overall performance of the model. On the other hand, it is sensitivite to noise because it is depend on the spectral form and fails to work in large datasets. PSO–ELM[3] enhances the performance in terms of high accuracy and also cost-effective. At the same time, the critical features extraction requires long time, and it is difficult to determine the initial weights. Phoneme Frequency Filtering Technique utilizes all training data to attain the maximum efficiency and also difficult to work analytically on the effect of a median filter. LSTM[4] provides robust performance even in noisy environment. Although, more memory is required to train the data and requires longer time to train the data. Feature Selection[5] reduces the model complexity and helps to train in faster way. This fails to work on real time applications. Back Propagation Technique has no parameters to tune apart from the numbers of input and is more flexible as it does not require prior knowledge about the network. The learning rate is too high, leads to unstable learning, and is quite sensitive to noisy data. Phone Recognition followed by language modeling (PRLM) reduces the classification error with an acceptable runtime in target languages. On the other side, this method provides low accuracy and decreases the speed in the performance. These challenges in the SLID encourage developing a new SLID model for better recognition and identification of spoken languages. The major contribution of the proposed spoken language feature selection and detection model is given below:

- To develop a new SLID model for regionally spoken languages of JK and Ladakhi with optimal feature selection and improved classification stage termed ISVM-RNN by adopting the AD-TSA for feature selection and detection with improved rate of accuracy.
- To present an effective optimal feature selection by AD-TSA for improving the classification performance with ISVM-RNN by minimizing the correlation among the features. This process focuses on attaining the most significant features for reducing the time and computational complexity.
- To propose an efficient classification SLID model termed ISVM-RNN with the cascaded hybrid classification by optimizing the parameters of SVM and RNN using AD-TSA with the aim of maximizing the accuracy and precision.
- To establish a novel AD-TSA algorithm for developing an optimal feature selection and classification techniques by selecting the optimal features and by optimizing the hidden neurons of RNN and the kernel functions in SVM to enhance the efficiency of the proposed SLID model.
- To validate the efficiency of the proposed SLID model on trained datasets using different performance metrics by comparing with existing algorithms and classifiers.

The remaining sections of this paper are given here. Section 2 discusses the proposed architecture including optimal feature selection using AD-TSA algorithm. Section 3 discusses results and discussion of the proposed implementation and the section 4 provides conclusion of this model.

Table 1. Features and challenges of existing spoken language identification using deep learning

Author [citation]	Methodology	Features	Challenges
Garain et al. [1]	DDMLP, DCNN and SSGAN	This helps to attain the precision to the maximum level.	If the separate classes other than the training data classes are used, then the performance will decrease because of the dependency on the parameter choice.
Deshwal et al. [2]	MFCC, PLP and RASTA-PLP	It provides high accuracy and minimum error in the overall performance of the model.	To work on large datasets. It is sensitive to noise because it is dependent on the spectral form.
Albadr et al. [3]	PSO–ELM	It enhances the performance in terms of high accuracy. It is cost-effective.	The critical features extraction requires long time. It is difficult to determine the initial weights.
Ma et al. [4]	LSTM	It provides robust performance even in noisy environment.	It requires longer time to train the data. More memory is required to train the data.
Das et al. [5]	Feature Selection	It reduces the model complexity and helps to train the data in much faster.	To work on real time applications.

2 Methodology

2.1 Spoken language identification using hybrid machine learning model

2.1.1 Proposed Architecture

The process of identifying the spoken language from a speech utterance is known as SLID. The automatic language identification from a speech sample is considered as a major challenge of SLID since it contains multi-lingual utterances. Nowadays, various language identification systems are used to identify the language through speech signal. However, the raw speech signal cannot be processed directly as it contains noise. The pre-processing should be done to the input signal to remove some unwanted distortions or noise for further efficient classification. Therefore, to get the accurate identification of languages, noise removal procedure is essential, which can be performed by adopting different techniques. The people speak more regional languages in JK. In this paper, we are focusing on languages spoken in JK and Ladakhi that are observed as Hindi, English, Dogri, Kashmir, Ladakhi, and Urdu. These languages have shared some common set of scripts and phonemes except English. Due to these similarities, it is very difficult to do the language identification process for such languages. Similarly, English has different accent or pronunciation based on people. This SLID can be implemented by several deep learning methods by using the voice signals from specific region. The acoustic modeling has used different features in the language identification process. However, the process of selecting optimal features was a major challenge in this modeling. One of the common deep learning methods is LSTM, which is implemented to identify the spoken languages. However, it was limited to process the short utterances of speech. The Deformable Neural Network (DNN) has used i-vectors based classification in the automatic language identification. However it offers low performance when deals with large amount of multi-lingual datasets. Therefore, improved deep learning with hybrid concept is used to solve the existing issues and to achieve high performance with better accuracy in the SLID process. The diagrammatic representation of the proposed SLID model is depicted in Figure 1.

The proposed SLID model aims to identify the spoken languages of JK and Ladakhi through improved feature selection and enhanced hybrid detection method. This proposed model involves different steps like "pre-processing, feature extraction, optimal feature selection, and detection or classification". Initially, the input speech signals that are gathered from different datasets are given to pre-processing phase for removing noise from the raw speech signals using SNG technique. The attained pre-processed signals are forwarded to feature extraction stage to minimize the redundant data without losing relevant information. The feature extraction process is done by using cepstral features such as MFCC and RASTA-PLP and spectral features such as spectral roll-off, spectral flatness, spectral centroid, spectral bandwidth, and spectral contrast. As the length of feature sets are more, the optimal features are selected from the above-extracted features by AD-TSA for enhancing the detection performance of proposed SLID model through solving the objective function with minimization of correlation. Here, the ISVM-RNN is proposed by optimizing the hidden neurons of RNN and kernel functions of SVM through AD-TSA by solving the objective function with maximization of accuracy and precision. The proposed ISVM-RNN is a cascaded hybrid classifier to handle 6 languages for improving the efficiency of classification of spoken languages of JK and Ladakhi.

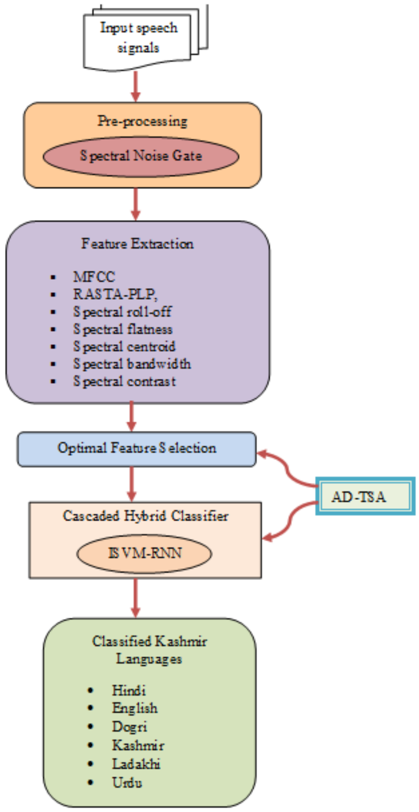


Fig 1. Architecture of Proposed SLID model

2.1.2 Dataset description

The proposed SLID model has gathered data from six different types of languages spoken in JK and Ladakh such as Hindi, English, Dogri, Kashmir, Ladakhi and Urdu. The dataset of English and Hindi are collected from standard corpora IIIT-H and VoxForge. The dataset of other languages are manually collected from different sources.

2.1.3 Pre-processing

The raw input speech signals has some unwanted distortions or noise thereby it cannot be processed directly through the SLID model. Hence, the input speech signal should undergo pre-processing method before the feature extraction process. The pre-processing method is used to analyze the input speech signals whether it is corrupted by some background noise. In this paper, the pre-processing of input speech signals S_{kl} is done by SNG technique.

SNG[(10)]: It is a common technique used for manipulation and resonant mix, which attenuates a signal corresponding to certain threshold. This technique is also used in general noise removal process. The sound spectrum has undergone some attenuation operations. The following steps follows the noise removal processing using SNG: (a) the noise audio clip is processed to calculate Fast Fourier Transform (FFT), (b) Statistics are calculated on FFT of the noise in terms of frequency, (c) By using the calculated statistics, a threshold value is calculated, (d) The FFT is calculated on the signal, (e) The signal FFT and the threshold are compared to determine a mask value, (f) The determined mask is smoothed through a filter over time and frequency, (g) Then, the smoothed mask is applied to the FFT of the signal and finally the pre-processed signals are attained by applying inverse FFT. Therefore, the pre-processed signals using SNG technique S_{kl}^{pr} are obtained, which is fit for further processes.

2.2 Optimal feature selection based on adaptive distance-based tunicate swarm algorithm

2.2.1 Feature Extraction Techniques

The pre-processed speech signals S_{kl}^{pr} are further given as input to the feature extraction process. The feature extraction is a process of deriving appropriate information from the speech utterances. This paper uses two types of feature extraction methods such as cepstral features and spectral features. The cepstral features include MFCC, RASTA-PLP and the spectral features such as spectral roll-off, spectral flatness, spectral centroid, spectral bandwidth, and spectral contrast.

Cepstral features**:** This technique is used to separate the speech signals into its source and its system components. The features are less correlated in the cepstral domain, and it is invariant to amplitude and transition changes.

MFCC[11]: It is a predominant feature extraction method for signifying the information of speech utterances. The continuous speech utterances are recognized by matching the input signal $S_{kl}^{pr}(c)$with a group of words or sentences. The first step is called as the parameterization. The input signal is transformed to the parameters to reduce the amount of redundant data. The familiar parameters in the recognition system are the MFCC. Let the input speech signal be represented as $S_{kl}^{pr}(c)$, and the evaluation techniques of the coefficients are described as below.

1) The calculation of energy spectrum is denoted on Eq (1)

$$\widetilde{z}(q) = \sum_{c=0}^{C_v-1} S_{kl}^{pr}(c)V(c)e^{-i2\pi qq}/C_v; \quad 0 \le q < C_v \tag{1}$$

In the above equation, the term C_v denotes the dimension of the Hanning window, and it corresponds to 30 ms. It is denoted in Eq. (2).

$$V(c) = \gamma_v\left(0.5 - 0.5\cos\left(\frac{2\pi c}{C_v - 1}\right)\right); \quad 0 \le c < C_v \tag{2}$$

Here, the term β_v refers to the normalization factor, such that the value of the root mean square of the window is unity. Eq. (3) defines the energy spectrum.

$$Z_q = |\widetilde{z}(q)|^2; \quad 0 \le q < Q \tag{3}$$

The term $Q$is assumed almost equal to $C_v/2$, since, half of the spectrum is considered.

2) The energy of each channel is computed in Eq (4)

$$H_j = \sum_{q=0}^{Q-1} \phi_i(q) Z_q; \quad 0 \le j < J \tag{4}$$

Here, the term J equals the triangular filters count ϕ_i and it is described using a constraint as in Eq. (5).

$$\sum_{q=0}^{Q-1} \phi_i(q) = 1; \quad \forall j \tag{5}$$

3) MFCC is calculated as in Eq (6)

$$FR_{dr}^{MFCC} = \gamma_d \sum_{j=0}^{J-1} \cos\left(l\frac{\pi}{I}(j+0.5)\right)\log_{10}(H_j) \tag{6}$$

The above equation can be described as scalar product among the log spectral energy vector and a vector of weighting factors WF_l as in Eq. (7).

$$WF_l = \left\{\cos\left(l\frac{\pi}{J}(j+0.5)\right) \quad 0 \le j < J\right\} \tag{7}$$

The MFCC is calculated in Eq. (8)

$$FR_{dr}^{MFCC} = \gamma_d \sum_{j=0}^{q-1} W\,F_{l,j}\log_{10}(H_j) \tag{8}$$

Here, the amplification factor is described by the term γ_d , the dynamic range of the coefficients m_c is indicated by FR_{dr}^{MFCC} . Thus, the MFCC extracted features are represented as FR_{dr}^{MFCC} , where $dr = 1, 2, \cdots DR$ and DR denotes the total number of MFCC extracted features that are attained as 3160.

RASTA-PLP[11]: The short term speech signals are represented by using Perceptual Linear Prediction (PLP) feature extraction method. RASTA-PLP is the enhanced form of PLP method, which overcomes the limitations of PLP technique.

This improved technique suppresses the adverse frequencies and increases the robustness of PLP in terms of noise. In RASTA-PLP method, critical band spectral resolution is applied for audible spectrum analysis, and band-pass filter is employed for smoothening spectral variations, which are performed in each-frequency sub-band. This process leads to derive a new spectral estimation, which is less prone to such variations. Then, the non-linear transformation process is done on the filtered speech signal spectral representation. The RASTA-PLP method tries to include the noise cancellation feature of the human auditory system and it is considered as the main advantage of this feature extraction method for the SLID system. The transformation function calculation of RASTA-PLP is formulated in Eq. (9)

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}\left(1 - 0.98z^{-1}\right)} \tag{9}$$

The fastest spectral change of the log spectrum is determined by the low cut-off frequency. The fastest spectral change is preserved in the output parameters, which is denoted by the high cut-off frequency. The higher values of the band-pass filter attenuate the convolution noise. The previous outputs are stored in the memory of RASTA filter. The current analysis results depend on this stored output values.

Thus, the RASTA-PLP extracted features are represented as FR_{dr}^{RASTA}, where the total number of RASTA-PLP extracted features are attained as 810.

Spectral Features**:** The spectral features are a kind of frequency-based features. They are commonly used to classify the speech and audio signal. It is obtained by transforming the time based signal into the frequency domain by means of FFT.

Spectral roll-off[12]: It is described as “the kl^{th}percentile of spectral distribution of the signal. The term kl ranges from 80% to 90%. It is the frequency attained below which the kl^{th} percentile of the magnitude distribution is focused”. Spectral roll-off calculation for the given input signal S_{kl}^{pr} is given in Eq. (10).

$$SRF = \textstyle\sum_{qr=0}^{mn_r} \left|S_{kl}^{pr}(qr)\right| = \frac{CR}{100} \displaystyle\sum_{qr=0}^{MN-1} \left|S_{kl}^{pr}(qr)\right| \tag{10}$$

Hence, the spectral roll-off extracted features are represented as FR_{dr}^{SPR} . The total number of spectral roll-off extracted features is attained as 158.

Spectral flatness[12]: The spectral flatness finds the differentiation among noise and harmonic like sounds. The spectral flatness is nearly zero for harmonic sounds and around one for noise like sounds. In power spectrum, it measures the uniformity in the frequency distribution. It is computed as ratio of the geometric mean to the arithmetic mean.

$$SFT = \frac{\textstyle\prod_{qr=0}^{MC-1} \mid S_{kl}^{pr}(qr)^{\frac{1}{MC}}}{\frac{1}{MC} \displaystyle\sum_{qr=0}^{MC-1} \left|S_{kl}^{pr}(qr)\right|} \tag{11}$$

Therefore, the spectral flatness extracted features are represented as FR_{dr}^{flat} . The total number of spectral flatness extracted features is attained as 158.

Spectral centroid[12]: It is “a measure utilized in the Digital Signal Processing (DSP) for characterizing a spectrum. It describes whether the center of mass of the spectrum is located. It is computed as the weighted mean of the frequencies available in the signal, and it is defined by means of a Fourier transform having their magnitudes as their weights” as in Eq. (12).

$$SCD = \frac{\sum_{qr=0}^{MC-1} fe(qr)mg(qr)}{\sum_{qr=0}^{MC-1} mg(qr)} \tag{12}$$

Here, the weighted frequency value or the magnitude is denoted by *mg(qr),* center frequency is denoted by. Hence, the spectral centroid extracted features are represented as FR_{dr}^{cen}. The total number of spectral centroid extracted features is attained as 158.

Spectral bandwidth[12]: In speech signal, “the spectral bandwidth is defined as the band width of signal at one-half the peak maximum used to determine the narrowness of a wave spectrum” and it is denoted in Eq. (13).

$$\varepsilon = \sqrt{1 - \frac{(mb_2 \cdot 2)}{mb_0 * mb_4}} \tag{13}$$

Hence, the spectral centroid extracted features are represented as FR_{dr}^{cen} centroid extracted features is attained as 158.

Spectral Contrast: It is defined as "the difference between the peak values and valley values of the spectrum," which is given here.

$$PK = \log\left(\frac{1}{\alpha PN}\sum_{ij=0}^{\alpha PN} S_{kl}^{pr}, PN+1\right) \tag{14}$$

$$VL = \log\left(\frac{1}{\alpha PN}\sum_{ij=0}^{\alpha PN} S_{kl}^{pr}, PN-ij+1\right) \tag{15}$$

Hence, the spectral contrast extracted features are represented as FR_{dr}^{cont}contrast extracted features is attained as 1106. Finally, the total number of extracted features is represented as FR_{dr}^{exfs}, where $FR_{dr}^{exxs} = \left\{FR_{dr}^{MFCC}, FR_{dr}^{RASTA}, FR_{dr}^{SPR}, FR_{dr}^{\text{flat}}, FR_{dr}^{cen}, FR_{dr}^{co\ nt}\right\}$ and it is given as input to the optimal feature selection process.

2.2.5 Optimal Feature Selection

Feature selection is a process of selecting optimal features from the set of features extracted from the feature extraction phase. The features selection process is differed from the feature extraction in such a way that the feature extraction process creates new features by using set of function or methods whereas the feature selection process delivers a sub-set of optimal features from the whole extracted features. In this paper, the features are selected from FR_{dr}^{exfs} by using AD-TSA for the effective process of classification. The optimal features obtained by using AD-TSA are represented as $FR_{of^*}^{opt}$, where *of** = 1,2, ...*OF* and *OF* denotes the total number of optimal features that are attained as 25. The classification performance of the proposed model is further enhanced by minimizing the correlation between optimal features.

$$\begin{matrix} obbj1_{fun}{=}argmin(corr) \\ \left\{FR_{of^*}^{opt}\right\} \end{matrix} \tag{16}$$

Correlation coefficients *corr* "consider the relative movements in the features and then define if there is any relationship between them".

$$co\ rr = \frac{OF\left(\Sigma FR_1^{\text{opt}}\, FR_2^{\text{opt}}\right) - \Sigma FR_1^{\text{opt}}\, \Sigma FR_2^{\text{opt}}}{\sqrt{\left[OF\Sigma\left(FR_1^{\text{opt}}\right)^2 - \left(\Sigma FR_1^{\text{opt}}\right)^2\right] O\,F\Sigma\left(FR_2^{\text{opt}}\right)^2 - \left(\Sigma FR_2^{\text{opt}}\right)^2\right]}} \tag{17}$$

Here, the term *OF* denotes first and second features, FR_1^{opt} and FR_2^{opt}denotes first and second features in the list respectively.

The diagrammatic representation of optimal feature selection is depicted in Figure 2.

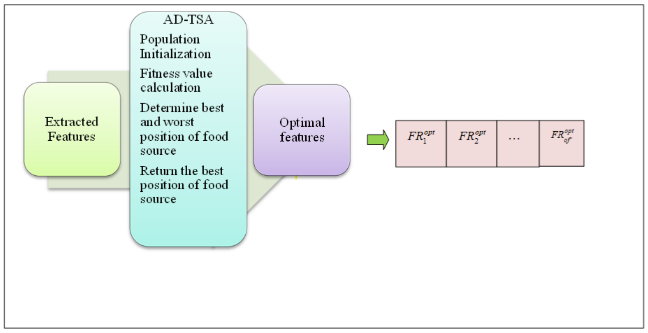

AD-TSA
Population Initialization
Fitness value calculation
Determine best and worst position of food source
Return the best position of food source
Extracted Features
Optimal features
FR_1^{opt}
FR_2^{opt}
...
FR_{of}^{opt}

Fig 2. Optimal feature selection method of proposed SLID model

2.3 Enhanced spoken language identification by improved support-vector-Recurrent neural network

2.3.1 Proposed ISVM-RNN for Spoken Language Identification

In general, the SVM and RNN techniques are widely used in the spoken language recognition. In the proposed model, a new approach named ISVM-RNN is established to improve the classification performance of SLID model and reduce the complexities of existing SVM and RNN. The proposed ISVM-RNN model is introduced by using AD-TSA through optimizing the kernel functions of SVM and hidden neurons of RNN. The selected optimal features $FR_{of^*}^{opt}$ are given as input to the ISVM-RNN for identification of languages spoken in JK and Ladakh. The classification is done based on the cascaded hybrid classification method, in which the input optimal features $FR_{of^*}^{opt}$ are given to SVM and RNN for 6 languages separately. The concept of hybrid classifier is based on taking the average of the classification score from both SVM and RNN. The average classification score is correlated with a threshold value assigned as 0.5. The proposed model uses a cascaded hybrid classifier to classify the languages under consideration. If the classification score is greater than 0.5, then the class is 1 (any language from cascaded structure), otherwise the class is 0. If the class is 0, then the features go to the next classifier to identify the rest languages. Similarly, all the other languages are classified based on the cascaded method.

SVM[13]: It is a supervised learning algorithm that can be applied for many regression and classification problems. In this paper, the SVM involves in efficient training and classification process of speech signals. SVM uses support vectors for the decision function, and it is formulated in Eq. (18).

$$S(y(t)) = \sum_{i=1}^{V} \delta_i^* x_i K_f \left(y_i^*, FR_{of^*}^{\text{opt}}(t) \right) + bs^* \tag{18}$$

Here, the term y_i^* denotes the i^{th} and vector of V support vectors, the class label is denoted by x_i and the term $FR_{of^*}^{opt}\ (t)$ indicates the t^{th} input frame vector, respectively. Lagrange multiplier δ^* and Optimization bias bs^* are obtained by solving a quadratic programming problem. The entire set of V support vectors is employed for a single input vector to generate an output of an SVM. The class of the input is determined by comparing the results of decision function with the predefined threshold. The kernel function of Radial Basis Function (RBF) is represented as $K_f\left(y_i^*, y(t)\right)$, , and it is formulated in Eq. (19).

$$K_f\left(y_i^*, y(t)\right) = \exp\left(-\beta \left\| y_i^* - y(t) \right\|^2\right) \tag{19}$$

Here, the kernel parameter is indicated by δ that is related with RBF width. The kernel function is one of the most frequently used functions for linearly-inseparable problems. Although SVM has provided superior performance in classification through feature extraction, the implementation is limited by its high memory requirement and computational intensity. Therefore, in the proposed SLID system, the kernel functions used in SVM are explained below, which are: (i) Linear kernel, (ii) Sigmoid kernel, (iii) Radial Basis Function (RBF) and (iv) Polynomial kernel, which are optimized using AD-TSA.

Linear kernel: It is a simple form of linear function. “It takes the inputs, multiplied by the weights for each neuron, and creates an output signal proportional to the input”.

Sigmoid kernel: This function is very simple, which takes “a real value as input and gives probability always between 0 and 1”. The main advantage is that, it provides good performance for classification.

Polynomial kernel: It represents ”the similarity of vectors (training samples) in a feature space over polynomials of the original features, allowing learning of non-linear models”.

RBF kernel: It is the most generalized form of kernelization and similar to Gaussian distribution “The RBF kernel function for two points computes the similarity or how close they are to each other”.

Recurrent Neural Networks (RNN)[14]: The process of RNN is based on a feed-back connection network, in which each and every node is linked with the other nodes (loops). The activation flow round in a loop is considered as the key feature of RNN that allows the network to do efficient training and temporal processing. The recurrent connections have MLP such as input layer, output layer and certain number of hidden layers. It also includes previous unit activation functions, feed back into the network along with the input. RNN use internal memory to process the sequence of input signals, which makes the network for the efficient process of speech recognition. The RNN takes high computational time due its recurrent nature, which needs further training. Therefore, it is further tuned by optimizing the hidden neurons

Finally, the efficiency of proposed model in terms of classification is enhanced by establishing an improved form of SVM and RNN, which is done by using AD-TSA. This process involves optimizing the kernel functions of SVM and hidden neurons of RNN. Therefore, the proposed ISVM-RNN classification method with cascaded hybrid nature enhances the performance of SLID model. The architecture of proposed ISVM-RNN classification model is depicted in Figure 3.

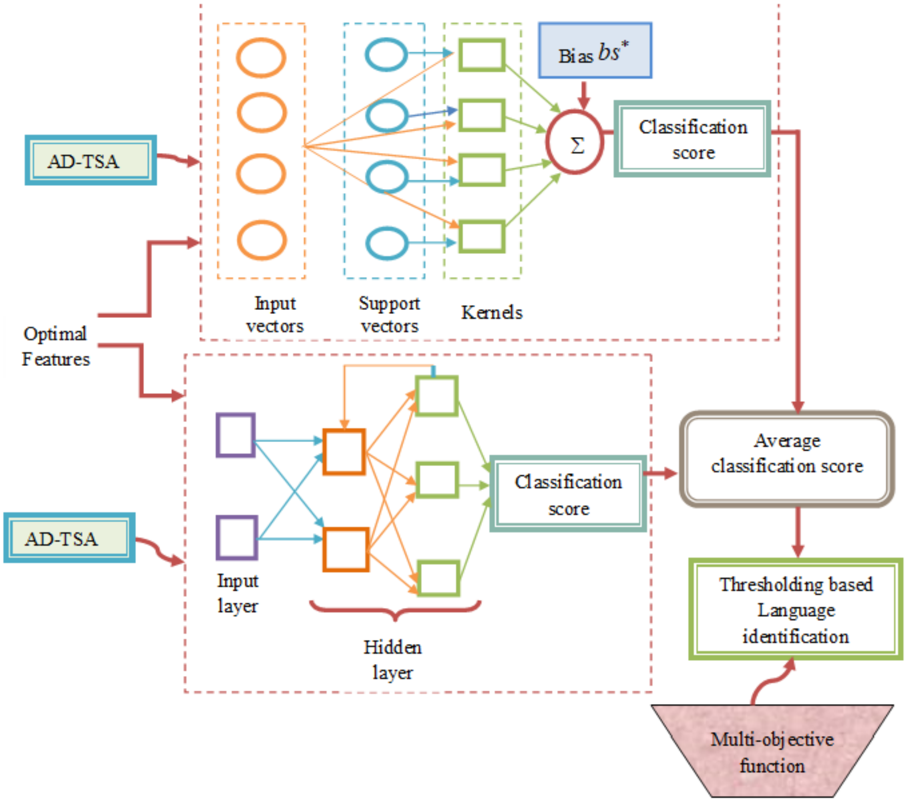


Fig 3. Achitecture of proposed ISVM-RNN classification model

2.3.2 Developed Objective Model

The main objective of proposed SLID model is to enhance the detection performance by solving the multi-objective function. The effective detection performance is done by using optimal feature selection through AD-TSA with the objective of minimizing the correlation among the features. Then, the classification phase uses the optimal features to improve the efficiency of the classification performance by proposing the ISVM-RNN through the optimization of kernel functions of SVM and hidden neurons of RNN, which is proved by maximization of accuracy and precision.

$$\mathrm{obj}\,2_{fu\,n} = \underset{\{k,mk,pk,rk,RNN_{hd}\}}{\arg\min}\left(\frac{1}{\mathrm{accy}+\mathrm{prcn}}\right) \tag{20}$$

Here, the activation functions like Linear kernel, Sigmoid kernel, polynomial kernel and RBF are represented as *lk, mk, pk* and *rk* respectively. Accuracy is referred as "the nearness of the measurements to a specific value". It is formulated in Eq. (21).

$$accr = \frac{(tr_{ps}+tr_{ng})}{(tr_{ps}+tr_{ng}+fl_{ps}+fl_{ng})} \tag{21}$$

Here, term tr_{ps} is denoted as the true positive, fl_{ps} is denoted as false positives, tr_{ng} true negatives and fl_{ng} is denoted as false negatives. Precision is referred as "the points that are stated to be positive especially it is used to declare what percentage of the points is truly positive" as denoted in Eq. (22).

$$prcn = \frac{tr_{ps}}{tr_{ps}+fl_{ps}} \tag{22}$$

Hence, the proposed ISVM-RNN efficiently classifies the languages spoken in JK and Ladakh.

2.3.3 Proposed AD-TSA

The detection of spoken language is effectively enhanced by the developed TSA algorithm namely AD-TSA. It focuses on selecting the optimal features and improving the detection process by introducing the new ISVM-RNN. It is further enhanced by optimizing the kernel functions of SVM and hidden neurons of RNN by using AD-TSA. In the proposed SLID model, the developed AD-TSA solves the search deflation issues in existing TSA and enhances the performance by providing optimal position of food source through the computation of best and worst solutions of search agents.

TSA(15): It is meta-heuristic algorithm, which has strong robustness, global optimization ability and fast convergence rate. It also applied to real-world optimization issues and it provides optimal mean variance and standard deviation values in any mathematical model of optimization techniques. Although, TSA provides fast convergence rate compared with existing algorithms, it faces local optima issues. In this paper, the existing issues of TSA are solved by developing a modified TSA algorithm namely AD-TSA. This AD-TSA improves the efficiency of the SLID system in terms of high competitive performance. The existing TSA use random parameter for updating the solutions, which lies in the range of [0,1]. In existing TSA, the search agents move towards the best neighbor by calculating the distance between the food source and the search agents. In our proposed TSA, the migration of search agents towards best neighbor is determined by the best position of food source and worst position of food source to attain the optimal solution. Moreover, the fitness value is calculated for all the solutions to validate the food source.

Among the many bio-inspired algorithms, TSA is considered as a special class of algorithm, which is based on the principle of jet propulsion and swarm behaviors during the searching and navigation process. TSA algorithm is inspired by using an insect namely tunicates, which lives in ocean. The tunicates has produced pale blue-green light, which is bright and can able to view from far distance due to its bio-luminescent capacity. Tunicates are able to find the location food source in the sea without knowing the search space, where the food source is available. The swarm behavior and jet propulsion are used in the process of searching the food source. The tunicates migrate to the depth of the ocean vertically by using the fluid jet-like propulsion. Hence, the improved optimization in the proposed model is achieved by using the major facts like jet propulsion and swarm behavior.

The mathematical modeling of jet propulsion is implemented by satisfying the three conditions such as migrate towards the position of best search agent, avoid the conflicts among search agents and remains close to the best search agent. Additionally, the locations of search agents are updated by the swarm behavior, which gives the best optimal solution. The position of new search agent is computed with the condition applied that the conflicts should be avoided between the other tunicates as formulated here with the applied vector $\overrightarrow{R}$.

$$\vec{R} = \frac{\vec{p}}{\vec{F}} \tag{23}$$

$$\vec{P} = a_2 + a_1 - \vec{M} \tag{24}$$

$$\vec{M} = 2 \cdot a_1 \tag{25}$$

In the above equations, the direction of water flow is indicated by the term $\overrightarrow{M}$ and the gravity force is denoted by the term $\overrightarrow{P}$. The random numbers are denoted by the variables a_1, a_2 and a_3 . The social force between the search agents is indicated by $\overrightarrow{F}$ and the equation is denoted in Eq. (26).

$$\vec{F} = [U_{\min} + a_1 \cdot U_{\max} - U_{\min}] \tag{26}$$

Here, the initial and subordinate speeds to build social interaction are represented by U_{min} and U_{max} , which has a range of 1 and 4, respectively. Next, the search agents are migrated towards the direction of best tunicate after avoiding the conflicts among the other tunicates and this process is formulated in Eq. (27).

$$\vec{U}V = \begin{cases} DE_{\text{best}} - nm_{\text{and}} \cdot \vec{U}_p(xy) & \text{if } FFT_{PR} > 0.5 \\ DE_{\text{wort}} - nm_{\text{and}} \cdot \vec{U}_p(xy) & \text{if } FFT_{PR} < 0.5 \end{cases} \tag{27}$$

Here, the term DE_{best} denotes the best position of food source and the term DE_{worst} indicates the worst position of food source regarding with the current position of search agents $\vec{U}_p(xy)$.

Here, the distance between the food source (best and worst position) and the search agents is referred by $\overrightarrow{U}V$, the current iteration is denoted by xy, the position of search agents is indicated by $\vec{U}_p(xy)$ and the random number of range [0,1] is denoted by n_{and}.

In the existing TSA model, the position of best neighbor is calculated by the distance among the food source and the search agent. In this proposed model, the position of best and worst food source is determined, which is formulated by assigning the random parameter FFT$_{PR}$. It is derived among the fitness value of all the solutions obj_{fun} and the mean of the fitness values $\mu\left(obj_{fi\,n}\right)$.

$$FFT_{PR} = \frac{obj_{fu\,n}}{\mu\left(obj_{fi\,n}\right)} \tag{28}$$

This random parameter helps to decide the best or worst food source for computing the distance among the food source and search agents.

The tunicate maintain its position converge towards the best search agent with food source. The vector $U_p\left(xy'\right)$ denotes the new position of tunicate concerned with position of food source $\vec{D}E$.

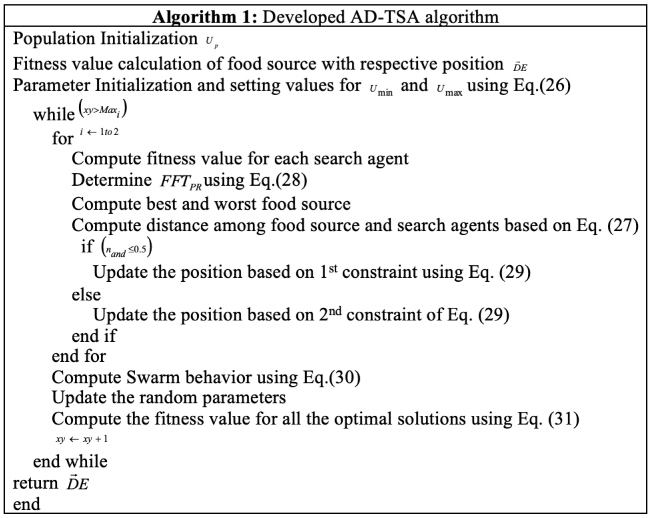
$$U_p\left(xy'\right) = \begin{cases} \vec{D}E + \vec{G}\cdot\vec{U}V & \text{if } n_{\text{and}} \geq 0.5 \\ \vec{D}E - \vec{G}\cdot\vec{U}V & \text{if } n_{\text{and}} < 0.5 \end{cases} \tag{29}$$

The mathematical model of swarm behavior of tunicate is denoted in Eq. (30). This process needs two optimal solutions like save and update the positions of the tunicate according to the position of near or best search agents.

$$\vec{U}_p(xy+1) = \frac{\vec{U}_p(xy)+\vec{U}_p(xy+1)}{2+a_1} \tag{30}$$

The proposed model is enhanced by the improved method of TSA through the new modification in swarm behavior and jet propulsion

The pseudo code of the AD- TSA algorithm is represented in Algorithm 1.



Algorithm 1: Developed AD-TSA algorithm

```
Population Initialization $U_p$
Fitness value calculation of food source with respective position $\vec{D}E$
Parameter Initialization and setting values for $U_{\min}$ and $U_{\max}$ using Eq.(26)
  while $(xy>Max_i)$
    for $i \leftarrow 1 to 2$
      Compute fitness value for each search agent
      Determine $FFT_{PR}$ using Eq.(28)
      Compute best and worst food source
      Compute distance among food source and search agents based on Eq. (27)
       if $(n_{and} \leq 0.5)$
         Update the position based on 1st constraint using Eq. (29)
      else
         Update the position based on 2nd constraint of Eq. (29)
      end if
    end for
    Compute Swarm behavior using Eq.(30)
    Update the random parameters
    Compute the fitness value for all the optimal solutions using Eq. (31)
    $xy \leftarrow xy + 1$
  end while
return $\vec{D}E$
end
```

Therefore, the developed AD-TSA algorithm provides better performance by improving the convergence rate and solves the existing TSA issues. The flowchart of the AD-TSA algorithm is represented in Figure 4.

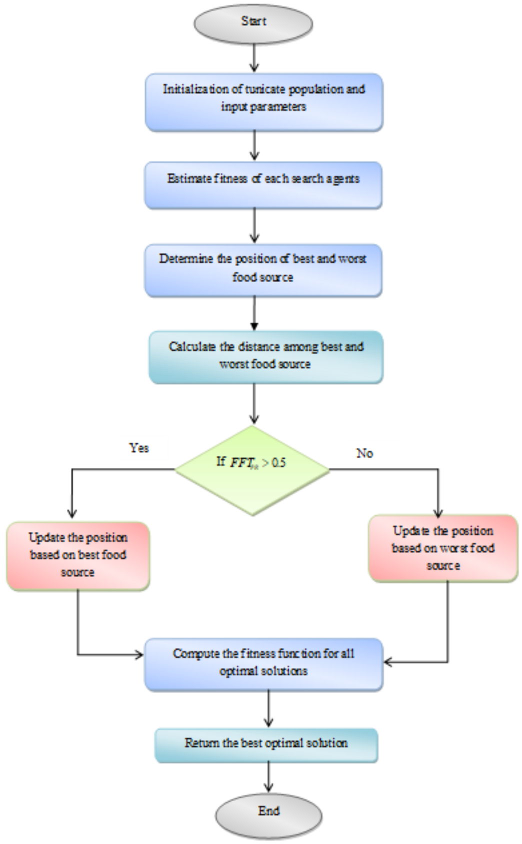

Start
Initialization of tunicate population and input parameters
Estimate fitness of each search agents
Determine the position of best and worst food source
Calculate the distance among best and worst food source
If $FFT_{tu} > 0.5$
Yes
No
Update the position based on best food source
Update the position based on worst food source
Compute the fitness function for all optimal solutions
Return the best optimal solution
End

Fig 4. Flowchart of the AD-TSA algorithm

3 Results and Discussion

3.1 Experimental setup

The developed SLID model focusing on languages spoken in JK and Ladakh was developed in Python. The performance assessment has been done between the conventional model and proposed model based on meta-heuristic algorithms and classifiers in terms of Type I and Type II measures. The "Type I measures are positive measures like Accuracy, Sensitivity, Specificity, Precision, Negative Predictive Value (NPV), F1Score and Mathews correlation coefficient (MCC), and Type II measures are negative measures like False positive rate (FPR), False negative rate (FNR), and False Discovery Rate (FDR)". The developed SLID model was analyzed with 10 number of population and 25 maximum numbers of iterations. The proposed meta-heuristic algorithm AD-TSA was evaluated with other algorithms such as "Particle Swarm Optimization (PSO)[16], Crow Search Optimization (CSO)[17], Moth flame optimization (MFO)[18] and TSA[15]", classifiers like BP[19], NN[20], SVM[21], RNN[14]. Since this the first attempt that TSA is used for language identification particulary for low resource languages that are spoken in Jammu & Kashmir along with Ladakhi language. As per the recent research in language identification there is not a single attempt that has worked on such a language group. Further, there is no work reported that used TSA for language identification. So we worked on this algorithm by improving TSA named AD-TSA. We compare our results with the PSO, CSO, MFO and TSA. Our improvised algorithm AD-TSA worked well in comparison to the available optimization algorithms.

3.2 Performance measures

The performance evaluation has considered various performance metrics that are given below:

The accuracy and precision is determined in Eq. (21) and Eq. (22)

(a) Sensitivity: It measures "the number of true positives, which are recognized exactly". (32)

(b) Specificity: It measures "the number of true negatives, which are determined precisely". (33)

(c) FPR: It is computed as "the ratio of count of false positive predictions to the entire count of negative predictions".

$$FPR = \frac{fl_{ps}}{fl_{ps} + tr_{ng}} \tag{34}$$

(d) FNR: It is "the proportion of positives which yield negative test outcomes with the test".

$$FNR = \frac{fl_{ng}}{tr_{n}g + tr_{ps}} \tag{35}$$

(e) NPV: It is the "probability that subjects with a negative screening test truly don't have the disease".

$$NPV = \frac{fl_{ng}}{fl_{ng} + tr_{ng}} \tag{36}$$

(f) FDR: It is "the number of false positives in all of the rejected hypotheses".

$$FDR = \frac{fl_{ps}}{fl_{ps} + tr_{ps}} \tag{37}$$

(g) F1 score: It is defined as the "harmonic mean between precision and recall. It is used as a statistical measure to rate performance".

$$F1 \text{ score } = \frac{2tr_{ps}}{(2\,(tr_{ps} + fl_{ps} + fl_{ng})} \tag{38}$$

(h) MCC: It is a "correlation coefficient computed by four values".

$$MCC = \frac{tr_{ps} \times tr_{ng} - fl_{ps} \times fl_{ng}}{\sqrt{(tr_{ps} + fl_{ps})\,(tr_{ps} + fl_{ng})\,(tr_{ng} + fl_{ps})\,(tr_{ng} + fl_{ng})}} \tag{39}$$

3.3 Performance analysis on meta heuristic-based algorithms

The AD-TSA algorithm of SLID model is compared with other meta-heuristic algorithms with different performance measures and learning percentages as shown in Figure 5 This diagram uses ten performance metrics like Accuracy, Sensitivity, Precision, FNR, FDR, FPR, NPV, MCC, F1-score to measure the performance of our proposed model AD-TSA-ISVM-RNN with the already available optimization algorithms like PSO, CSO, MFO and TSA by combing with the improved SVM-RNN classifer. The AD-TSA is performed 0.20% better than PSO, 0.10% better than CSO, 0.31% better than MFO and 0.15% better than TSA in terms of accuracy, for the learning percentage 65. Likewise, the proposed model reaches high percentage values when compared with the existing algorithms for all the other performance metrics while various learning percentages. The developed model reaches peak values for most of the learning percentages. Suddenly, there is a deviation in the value while varying learning percentages. For example, while considering the learning percentages between 50 and 55 in terms of sensitivity, the proposed model deviates from 99.55 to 99.45 and back to the peak value at the learning percentage 60. Although, there is some deviations for a specific learning percentage, the proposed model never lose its consistency and consequently provides better performance for the maximum numbers of learning percentages. Therefore, the results of graphical representation shows that the proposed AD-TSA algorithm-based SLID model delivers superior performance than the existing algorithms. Since there is no a single work related with the TSA used in the process of classification of languages particularly for the language spoken in Jammu & Kashmir along with the Ladakhi language. The reason may the due to non-availability of speech corpuses of these low resource languages. We have compare our results with the existing optimization techniques and it showed that our proposed model performed well.

3.4 Performance analysis on existing classifiers

The proposed ISVM-RNN classification technique of SLID model is compared with the other classifiers with various performance metrics like Accuracy, Sensitivity, Precision, FNR, FDR, FPR, NPV, MCC, F1-score and learning percentages, and it is represented in . The proposed model AD- ISVM-RNN is compared with the classifiers like NN, BP, SVM and RNN. In terms of FNR, the proposed ISVM-RNN classification technique is attained 90% better than back propagation technique, 90.9%

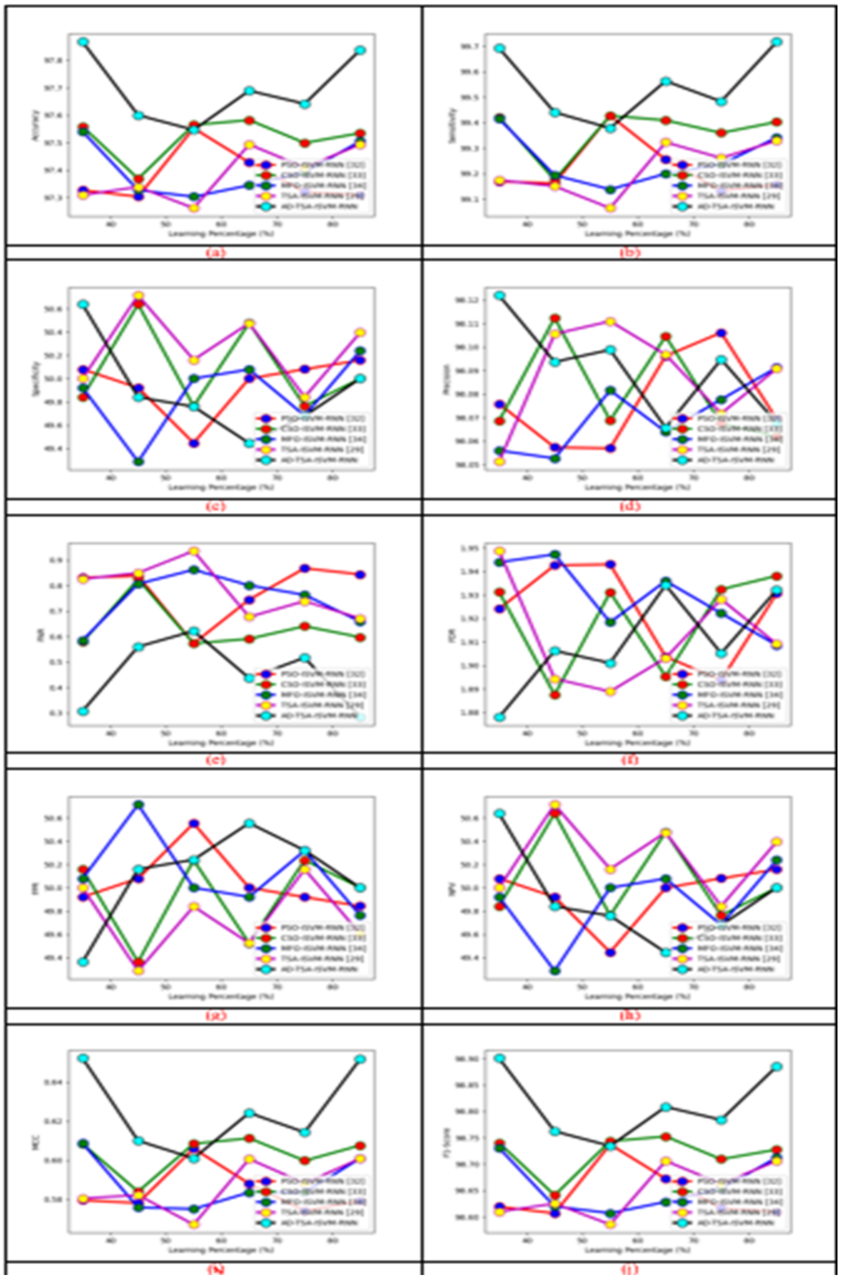


Fig 5. Performance ananlysis of the proposed SLID model on meta-heuristc algorithms in terms of "(a)Accuracy, (b)Sensitivity, (c)Specifity, (d)Precision, (e)FNR, (f)FDR, (g)FPR, (h)NPV, (i)MCC, (j) F1-score"

better than DCNN, 88.88% better than SVM and 91.8% better than RNN for the learning percentage 35. Similarly, the proposed ISVM-RNN provides enhanced performance for all the performance measures while changing the learning percentages. There may some sudden deviations for a certain learning percentage. For example, while considering NPV, the proposed ISVM-RNN provides less error rate except for the learning percentage 35. Even though, the proposed ISVM-RNN shows deviations for some learning percentage, it provides better performance for maximum number of learning percentages in different performance measures. Therefore, the results performance analysis shows that the proposed ISVM-RNN based SLID model offers improved performance than the existing classifiers.

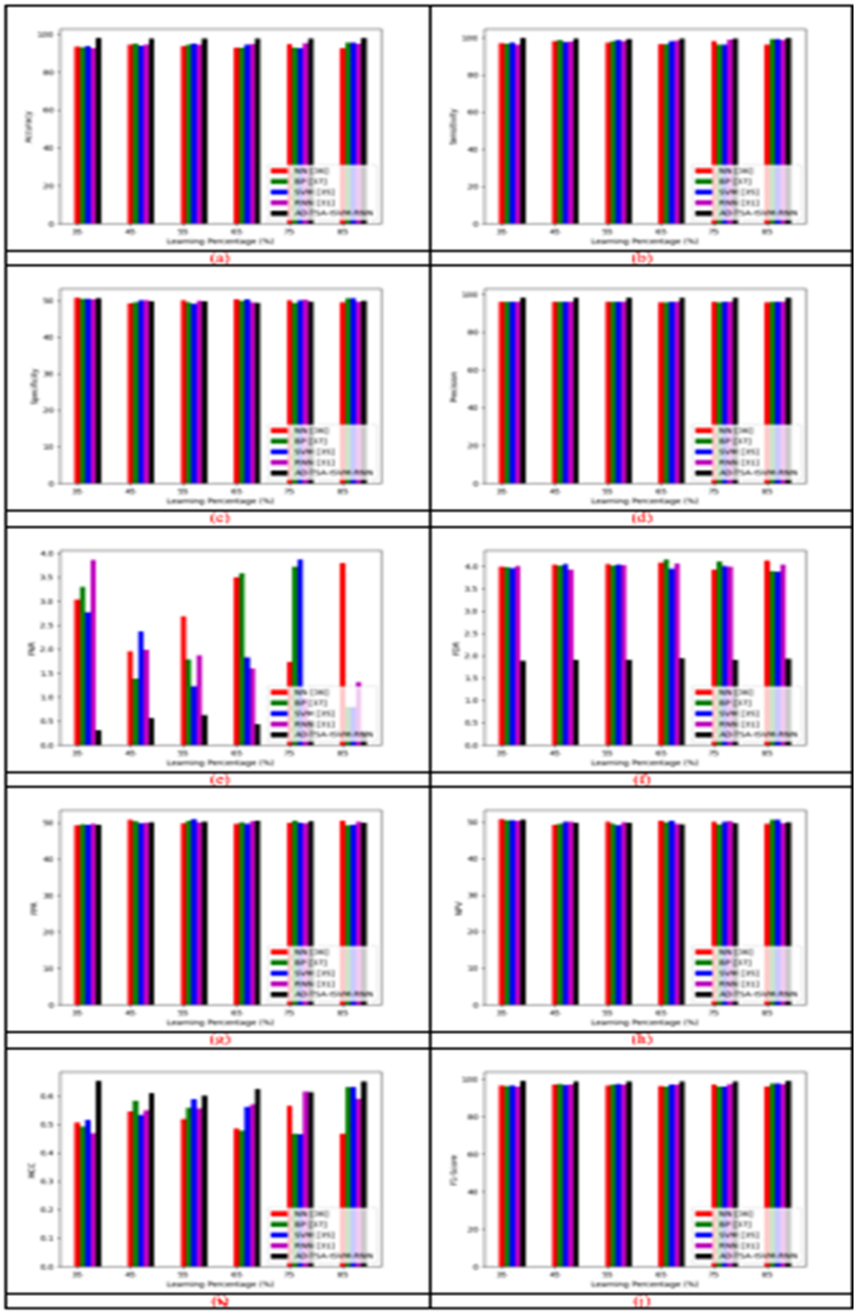


Fig 6. Performance ananlysis of the proposed SLID on classifiers model in terms of "(a)Accuracy, (b)Sensitivity, (c)Specifity, (d)Precision, (e)FNR, (f)FDR, (g)FPR, (h)NPV, (i)MCC, (j) F1-score"

3.5 Overall performance analysis

In order to give the overall performance analysis of our proposed model and to make comparison with the already state-of-the-art, there is not a single work that has used TSA as an optimization technique to reduce the feature set and there is not a single work related with the Ladakhi language. So we worked on such a diversified language set that uses languages of Jammu & Kashmir along with the Ladakhi language. But, as the performance report the overall performance analysis on the proposed SLID model is represented in Tables 2 and 3. In Table 2, the sensitivity of the proposed SLID is 0.3% better than PSO, 0.12% better than CSO, 0.24% better than MFO, and 0.22% better than TSA. Likewise, in Table 3, the FDR of the proposed SLID model is 51.45% better than back propagation 53.57% better than DCNN, 52.38% better than SVM and 52.29% better than RNN. Similarly, for all the performance measures the proposed SLID based on developed TSA algorithm and SV-RNN model provides better performance than the existing meta-heuristic algorithms and classifiers. Therefore, the overall analysis shows that the proposed SLID model offers better recognition and classification performance than the conventional models.

Table 2. Overall performance analysis of the proposed slid model with meta-heuristic based algorithms

Algorithms	PSO-ISVM-RNN [16]	CSO-ISVM-RNN [17]	MFO-ISVM-RNN [18]	TSA-ISVM-RNN [15]	AD-TSA-ISVM-RNN
Accuracy	0.973215	0.974993	0.973926	0.974104	0.976415
Sensitivity	0.991324	0.993597	0.992368	0.992612	0.994831
Specificity	0.500803	0.49763	0.496815	0.498418	0.496795
Precision	0.981062	0.980676	0.980777	0.980718	0.980948
FPR	0.499197	0.50237	0.503185	0.501582	0.503205
FNR	0.008676	0.006403	0.007632	0.007388	0.005169
NPV	0.500803	0.49763	0.496815	0.498418	0.496795
FDR	0.018938	0.019324	0.019223	0.019282	0.019052
F1-Score	0.986166	0.987094	0.986539	0.986629	0.987841
MCC	0.574133	0.599839	0.583656	0.588307	0.614366

Table 3. Overall performance analysis of the propsoed slid model with existing classifiers

Classifiers	BP [19]	NN [20]	SVM [21]	RNN [14]	AD-TSA-ISVM-RNN
Accuracy	0.946844	0.927526	0.927111	0.95283	0.976415
Sensitivity	0.982713	0.96283	0.961337	0.990119	0.994831
Specificity	0.500796	0.494099	0.500399	0.501939	0.496795
Precision	0.960754	0.958958	0.959985	0.96006	0.980948
FPR	0.499204	0.505901	0.499601	0.498061	0.503205
FNR	0.017287	0.03717	0.038663	0.009881	0.005169
NPV	0.500796	0.494099	0.500399	0.501939	0.496795
FDR	0.039246	0.041042	0.040015	0.03994	0.019052
F1-Score	0.971609	0.96089	0.96066	0.974858	0.987841
MCC	0.565083	0.467749	0.465516	0.614657	0.614366

4 Conclusion

In this work, a novel approach of SLID for spoken languages of JK and Ladakh was introduced. The input speech signals of languages under consideration languages were given as input to the SNG for pre-processing the signals. Then, the feature extraction process was done by cepstral and spectral techniques. The optimal features were selected by the ISVM-RNN, which is optimized using AD-TSA. The optimized ISVM-RNN has finally identified the languages of spoken in JK and Ladakh using the optimal features. The overall analysis on the proposed AD-TSA-ISVM-RNN has provided promising results over conventional meta- heuristic algorithms. While considering accuracy in the overall analysis, the proposed AD-TSA-ISVM-RNN has achieved 3.12% better than BP, 5.27% better than NN, 5.32% better than SVM, and 2.47% better than RNN. Therefore, it was concluded that the proposed AD-TSA-ISVM-RNN model has achieved superior performance in detection of languages spoken in JK and

Ladakh. In the future, the proposed model can be made more robust by adding more speech samples from different speakers and incorporating different accents for the same language, and extends this model via new innovative and well-performing deep learning models.

References

1) Garain A, Singh PK, Sarkar R. FuzzyGCP: A deep learning architecture for automatic spoken language identification from speech signals. *Expert Systems with Applications*. 2021;168(114416):114416. Available from: https://doi.org/10.1016/j.eswa.2020.114416.
2) Deshwal D, Sangwan P, Kumar D. A Language Identification System using Hybrid Features and Back-Propagation Neural Network. *Applied Acoustics*. 2020;164(107289):107289. Available from: https://doi.org/10.1016/j.apacoust.2020.107289.
3) Albadr MAA, Tiun S. Spoken Language Identification Based on Particle Swarm Optimisation–Extreme Learning Machine Approach. *Circuits, Systems, and Signal Processing*. 2020;39(9):4596–4622. Available from: https://doi.org/10.1007/s00034-020-01388-9.
4) Ma Z, Yu H, Chen W, Guo J. Short Utterance Based Speech Language Identification in Intelligent Vehicles With Time-Scale Modifications and Deep Bottleneck Features. *IEEE Transactions on Vehicular Technology*. 2019;68(1):121–128. Available from: https://doi.org/10.1109/TVT.2018.2879361.
5) Das A, Guha S, Singh PK, Ahmadian A, Senu N, Sarkar R. A Hybrid Meta-Heuristic Feature Selection Method for Identification of Indian Spoken Languages From Audio Signals. *IEEE Access*. 2020;8:181432–181449. Available from: https://doi.org/10.1109/ACCESS.2020.3028241.
6) Bakshi A, Kopparapu SK. Feature selection for improving Indian spoken language identification in utterance duration mismatch condition. *Bulletin of Electrical Engineering and Informatics*. 2021;10(5):2578–2587. Available from: https://doi.org/10.11591/eei.v10i5.3173.
7) Alashban AA, Qamhan MA, Meftah AH, Alotaibi YA. Spoken Language Identification System Using Convolutional Recurrent Neural Network. *Applied Sciences*. 2022;12(18):9181. Available from: https://doi.org/10.11591/eei.v10i5.3173.
8) Singh G, Sharma S, Kumar V, Kaur M, Baz M, Masud M. Spoken language identification using deep learning. 2021. Available from: https://doi:10.1155/2021/5123671.
9) Thukroo IA, Bashir R. Spoken Language Identification System for Kashmiri and Related Languages Using Mel-Spectrograms and Deep Learning Approach. *2021 7th International Conference on Signal Processing and Communication (ICSC)*. 2021;p. 250–255. Available from: https://doi.org/10.3390/app12189181.
10) Hou JC, Wang SS, Lai YH, Tsao Y, Chang HW, Wang HM. Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*. 2018;2(2):117–128. Available from: https://doi.org/10.1109/TETCI.2017.2784878.
11) Deshwal D, Sangwan P, Kumar D. A Language Identification System using Hybrid Features and Back-Propagation Neural Network. *Applied Acoustics*. 2020;164:107289. Available from: https://doi.org/10.1016/j.apacoust.2020.107289.
12) Sharma G, Umapathy K, Krishnan S. Trends in audio signal feature extraction methods. *Applied Acoustics*. 2020;158:107020. Available from: https://doi.org/10.1016/j.apacoust.2019.107020.
13) Lim C, Lee SR, Chang JH. Efficient implementation of an SVM-based speech/music classifier by enhancing temporal locality in support vector references. *IEEE Transactions on Consumer Electronics*. 2012;58(3):898–904. Available from: https://doi.org/10.1109/TCE.2012.6311334.
14) Dudhrejia H, Shah S. Speech Recognition using Neural Networks. *International Journal Of Engineering Research & Technology*. 2018;7. Available from: https://www.ijert.org/research/speech-recognition-using-neural-networks-IJERTV7IS100087.pdf.
15) Satnamkaur LK, Awasthi AL, Sangal G. Tunicate Swarm Algorithm: A new bio-inspired based metaheuristic paradigm for global optimization. *Engineering Applications of Artificial Intelligence*. 2020;90. Available from: https://doi.org/10.1016/j.engappai.2020.103541.
16) Selvaraj L, Ganesan B. Enhancing Speech Recognition Using Improved Particle Swarm Optimization Based Hidden Markov Model. *The Scientific World Journal*. 2014;2014:1–10. Available from: https://doi.org/10.1155/2014/270576.
17) Ouadfel S, Elaziz MA. Enhanced Crow Search Algorithm for Feature Selection. *Expert Systems with Applications*. 2020;159:113572. Available from: https://doi.org/10.1016/j.eswa.2020.113572.
18) Sapre S, S M. Emulous mechanism based multi-objective moth–flame optimization algorithm. *Journal of Parallel and Distributed Computing*. 2021;150:15–33. Available from: https://doi.org/10.1016/j.jpdc.2020.12.010.
19) Deshwal D, Sangwan P, Kumar D. A Language Identification System using Hybrid Features and Back-Propagation Neural Network. *Applied Acoustics*. 2020;164:107289. Available from: https://doi.org/10.1016/j.apacoust.2020.107289.
20) Hou JC, Wang SS, Lai YH, Tsao Y, Chang HW, Wang HM. Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*. 2018;2(2):117–128. Available from: https://doi.org/10.1109/TETCI.2017.2784878.
21) Chau G, Kemper G. One Channel Subvocal Speech Phrases Recognition Using Cumulative Residual Entropy and Support Vector Machines. *IEEE Latin America Transactions*. 2015;13(7):2135–2143. Available from: https://doi.org/10.1109/TLA.2015.7273769.