International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
IJISAE
ISSN:2147-6799 www.ijisae.org Original Research Paper

Automatic Speech Recognition System using Bio-Inspired Optimization and Convolutional Neural Network

[1]K. Pavan Raju, [2]Dr. A. Sri Krishna, [3]Dr. M. Murali

Submitted: 03/02/2024 **Revised**: 11/03/2024 **Accepted**: 17/03/2024

Abstract: The incorporation of machine learning methods has led to significant breakthroughs in Automatic Speech Recognition (ASR) systems. This study presents a new method that integrates the Opposition Whale Optimization algorithm (OWOA) with Convolutional Neural Network (CNN) to improve the efficiency of ASR systems. The Opposition Whale Optimization method, which draws inspiration from the social behavior of humpback whales, provides a distinct capacity to explore and exploit in order to optimize the parameters of the ASR system. The method employs opposition-based learning to achieve a well-rounded exploration of the search space, resulting in improved convergence speed and prevention of premature convergence. The use of OWOA, in combination with a CNN architecture, allows for the extraction of hierarchical characteristics from voice data. CNNs have achieved notable success in several pattern recognition applications owing to their capacity to grasp spatial and temporal connections in data. By using the capabilities of CNNs, the suggested ASR system may efficiently acquire distinctive characteristics from unprocessed audio data, resulting in enhanced accuracy in recognizing speech. The suggested technique is validated by conducting experimental assessments on typical speech datasets. The performance increases produced by integrating OWOA and CNN are assessed by conducting comparative assessments against baseline ASR systems. The findings indicate that the suggested ASR system surpasses current approaches in terms of accuracy in recognizing speech and speed of convergence, hence highlighting its potential for practical applications.

***Keywords**: Automatic Speech Recognition, Opposition Whale Optimization, Convolutional Neural Network, Machine Learning, Optimization Algorithms.*

1. Introduction

ASR systems have become essential elements in a wide range of applications, including virtual assistants, dictation systems, and voice-controlled devices. The efficacy of ASR systems is greatly contingent upon the precision of speech recognition, which is, in turn, contingent upon the resilience of the underlying algorithms and models. Machine learning methods have been used to make great progress in developing ASR technology throughout the years.

CNNs have been prominent in recent years as very effective techniques for extracting significant characteristics from unprocessed data in many fields, such as computer vision and natural language processing. CNNs have shown exceptional efficacy in capturing spatial and temporal relationships in data, making them highly suitable for tasks that need pattern detection and classification. Within the field of ASR, CNNs have shown significant potential in acquiring distinguishing characteristics from spectrograms or other forms of speech signal representations.

In contrast, CNN achieves translational invariance by using fewer parameters and duplicating weights across time and frequency. Deep neural networks (DNNs) are not affected by the arrangement of data and may process it in any predetermined sequence without affecting the network's output. This means that DNNs are not influenced by the input topology [1]. The spectral representations of speech have strong correlations, and using CNNs to simulate these local correlations has been shown to be beneficial in several fields [2]. The majority of CNN image recognition research use a limited number of convolutional layers followed by fully linked layers.

The increased computer power, greater availability of training data, and improved software engineering have all contributed to the remarkable successes of deep neural networks, as well as advancements in learning processes. Prior to implementing the discriminative model, the use of a generative backpropagation learning approach, specifically a layer-by-layer pre-training strategy, was responsible for the initial success in modifying the weights for acoustic modeling. However, subsequent research has revealed that pre-training the generative model is not suitable when there is a

[1]*Research Scholar, Department of CSE, Centurion University of Technology and Management, Andhra Pradesh, India.*
Pavanraju6999@gmail.com
[2]*Associate Professor, Shri Vishnu Engineering College for Women (A), Department of Information Technology, Vishnupur, Bhimavaram, Andhra Pradesh, India.*
Srikrishna.au@gmail.com
[3]*Professor, Department of Electronics and Communication Engineering, Centurion University of Technology and Management, Andhra Pradesh, India.*
muralitejas@cutmap.ac.in

substantial amount of labeled data available. Backpropagation may be started using randomly assigned weights, as long as their ranges are well defined to prevent excessively large or tiny initial error derivatives.

2. Literature Survey

Dua et al., in [3] examine the efficacy of CNNs in the field of voice recognition, with a particular focus on their ability to process tonal speech signals. The study expands upon CNN-based methods to tackle less prevalent tone speech using a specially created database. The CNN architecture employs TensorFlow and Praat for segmentation, using six layers. The MFCC approach is used to extract features, including both speech and background music characteristics. The results show the exceptional performance of the approach based on CNN.

Jagannath et al., in [4] have observed that advancements in communications technology have facilitated the incorporation of cognitive computations into smart healthcare systems. This integration has resulted in improved patient treatment by enabling the exchange of information and boosting scalability. Remote recording enhances the identification of respiratory diseases via speech, providing a rapid and noninvasive approach. Nevertheless, achieving a trade-off between achieving high levels of accuracy and keeping computing complexity low is a significant challenge. This research presents a neural network with minimal complexity that utilizes a cascaded PFLANN to improve classification accuracy without sacrificing efficiency. The evaluation conducted across several respiratory disorders reveals that the model exhibits excellent performance in terms of accuracy and complexity.

Tusar et al., in [5] Early detection of COVID-19 presents challenges owing to its fast transmission and widespread public apprehension. Speech-based detection provides a secure approach by using speech data that may be readily recorded. The study of Mel Frequency Cepstral Coefficient (MFCC) is a powerful instrument for this goal, but its effectiveness depends on the conversion of frequency scale and the range of filters used. Conventionally, fixed values are used, however, speech signal characteristics exhibit variations across different disorders. The primary emphasis of COVID-19 detection lies in analyzing coughing noises, which are distinguishable from normal speech signals. Enhancing the frequency range and refining the conversion scale enhances the efficacy of detection. Speech augmentation is performed before feature extraction in order to improve accuracy. This research presents the COVID-19 Coefficient (C-19CC) by optimizing these parameters. Comparative analysis assesses the effectiveness of these aspects.

Rajni Sobti et al., in [6] have made important advancements in the field of voice recognition technology, which have improved the way humans and machines communicate with each other. Nevertheless, this advancement has mostly benefited widely spoken languages, disregarding the linguistic variety of less-resourced languages. Speech recognition in these languages encounters difficulties as a result of restricted linguistic resources and data. This article aims to fill this need by specifically concentrating on the development of ASR systems tailored for children who speak Punjabi. The research obtained and divided speech samples using PRAAT, then transcribed and extracted features based on MFCC. Various models were used in acoustic modeling, ultimately resulting in the implementation of a DNN-HMM model to improve accuracy. The results indicate that the proposed DNN-HMM model has achieved a higher accuracy rate of 83.9% compared to current techniques.

The focus of ASR system development in India has mostly been on English, Hindi, and Marathi, out of the 22 primary languages in the country. Consequently, there is inadequate Automatic Speech Recognition (ASR) capability for languages other than Punjabi. Kadyan (in [7]) has noted a relative scarcity of research on the advancement of Automatic Speech Recognition (ASR) in Punjabi, in comparison to languages like English and Italian.

Hasija et al. (8) found that the effectiveness of ASR systems is heavily influenced by the quantity and quality of the training data. However, the scarcity of data and the variability in children's speech might have a detrimental effect on the accuracy of automated speech recognition (ASR) systems. This is particularly evident in languages like Punjabi, which display tonal features and have restricted resources. This research aims to address this challenge by examining the feasibility of Automatic Speech Recognition (ASR) performance via the assessment of two separate corpora. The objective of using Tacotron, a synthetic speech generation system for Punjabi, is to overcome the limitations caused by the scarcity of available data. Tacotron's synthetic speech audios are combined with existing data sets and assessed on a Punjabi children's Automatic Speech Recognition (ASR) system that uses Mel Frequency Cepstral Coefficients (MFCC) + pitch feature extraction and DNN acoustic modeling. The merged collection of texts exhibits a decreased Word Error Rate (WER) in the Automatic Speech Recognition (ASR) system, with a Relative Improvement (RI) ranging from 9% to 12%.

The objective of the work is given below:

1. The justification for employing the deep learning approach DBN for feature extraction and the metaheuristic algorithm whale optimization algorithm (WOA) to optimize the output of the DNN classifier is explained.
2. The CNN and its architecture are described also OWOA is a contemporary nature-inspired technique is employed. This meta-heuristic optimization replicates the behavioral patterns of humpback whales.
3. The experimental evaluation approach, which comprises numerous performance evaluation criteria, is outlined.

3. Proposed Work

The topology of the deep neural network (DNN) is modified by optimizing the hidden layers (HL) and the neurons inside the hidden layers using an opposition-based whale optimization algorithm (OWOA). The objective is to enhance the precision of identifying continuous speech signals in real-time while simultaneously reducing the processing time.

3.1 Convolutional Neural Network

CNN refers to a collection of deep learning neural networks. A CNN consists of the following layers: The layers of a convolutional neural network typically consist of a convolutional layer, a rectified linear unit (ReLU) layer, a pooling layer, and a fully connected layer. Localness, weight sharing, and pooling are three key characteristics of CNN. Each of them have the capacity to enhance speech recognition. The spatial arrangement of the convolution layer's units enhances resistance to non-white noise by allowing some frequency bands to remain unaffected while others get corrupted. As a result, the algorithm may use untouched regions to calculate favorable attributes on a small scale. Only a minuscule subset of characteristics is impacted by the spectrum and the disturbance. Weight sharing enhances model resilience and mitigates overfitting by including information from many frequency bands in the input, rather than relying on a single location.

Furthermore, weight sharing reduces the number of weights that must be obtained in the network. To have the property of pooling, it is required to share both location and weight. Pooling involves the aggregation of feature values that have been computed at many places, resulting in a single representative value. The pooling approach is able to catch small changes in the features, particularly when max-pooling is used, even when the input patterns are moved significantly along the frequency axis. This is especially beneficial for dealing with minor frequency variances that sometimes occur in speech broadcasts.

Architecture

From a local perspective, time is not a pressing problem. A solitary input window for the CNN, like to earlier DNNs used for speech, would include a substantial amount of context, often ranging from 9 to 15 frames. The conventional use of MFCCs is limited by a significant drawback in terms of frequency. This issue arises from the fact that the Discrete Cosine Transform (DCT) reassigns the spectral energies to a different basis, potentially compromising their spatial relationship.

The convolutional and pooling layers, while generating input feature charts, train their internal components using appropriate methodologies. Diagrams may be used to organize the components of the convolution and pooling layers, which are comparable to those of the input layer. In CNN nomenclature, a sequence of convolution and pooling layers is sometimes referred to as a single "layer" of CNN. Therefore, a deep Convolutional Neural Network (CNN) is composed of two or more individuals organized in a certain order. Figure 1 presents a detailed representation of the complex architecture of the CNN.

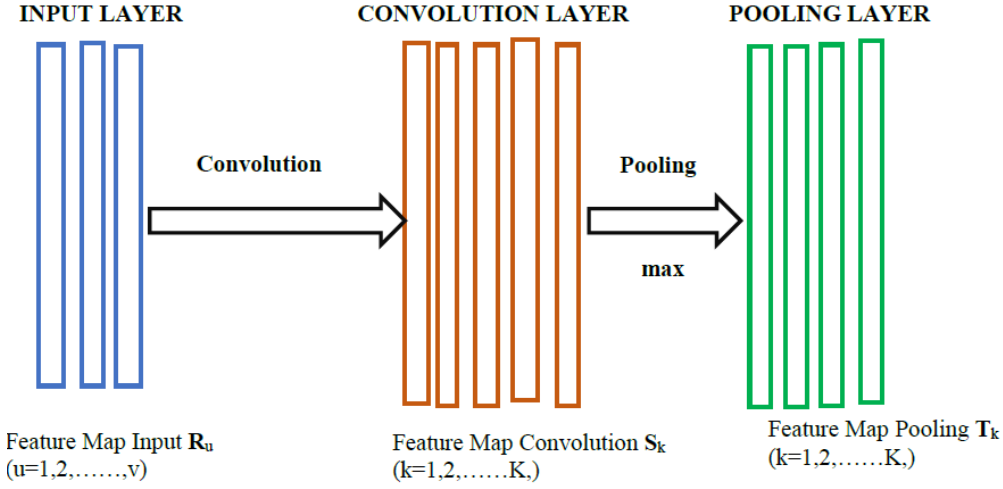


Fig 1: Architecture of CNN

The convolution layer unit for limited sharing of weight can be calculated as:

$$Z_{0,p,q} = \gamma \sum_{c=1}^{p} \sum_{l=1}^{q} \chi_c(o-1)Xe + l + q - 1W_{o,c,l} + W_{o.x,j} \quad \square \qquad (1)$$

The variable wo,,, represents the lth mapping of convolution weight from the cth input feature map to the pth convolution map in the oth segment. The variable q spans from 1 to H, which represents the pooling size.

$$n_{i,j} = \sum_{\mathrm{q}=1}^{\mathrm{H}} \mathrm{Z}_{\mathrm{o,p,q}} \qquad (2)$$

where, H represents the pooling size and s is the shift size

The output layer in CNN is

$$p_{i,m} = k \sum_{q=1}^{H} Z_o(m-1)XS + q \qquad (3)$$

The scaling factor k is a variable that may be acquired via learning. The picture undergoes detection with the condition that the pooling size matches the shift size. It has been shown that max-pooling yields better results than average-pooling when the assembling windows are non-overlapping and have no gaps between them.

3.2 Opposition Whale Optimization Algorithm

The study of machine learning relies heavily on three main processes: learning, optimization, and search. Algorithms acquire knowledge from past data or instructions, optimize approximated answers, and search for a solution in vast domains. There is a wide variety of problems, and algorithms are based on many different kinds of biological, mental, and environmental occurrences. Learning often starts at an inopportune moment. We begin our process by starting at the foundation and progressing towards an already established solution. A few examples are the random selection of evolutionary algorithms' parameter populations, the unpredictable determination of reinforcement agents' active policies, and the random setting of neural net weights. If the random estimate is somewhat close to the optimal solution, it may result in rapid convergence.

Hence, it is reasonable to observe that if we commence with a completely arbitrary conjecture that is significantly far from the actual solution, for instance, in the most unfavorable scenario, at the opposite end, the process of approximation, exploration, or optimization will need a much longer duration, or in the most unfavorable scenario, will become very difficult to manage. Without previous information, we will not make any remarkable first assumptions. Consequently, we need to be observing in every direction simultaneously, or, to be more precise, in the opposite way. The first step in finding the opposite number X' is essential while seeking Xi and acknowledging the potential benefits of searching in the other direction.

$$X' = a + b - X_i \qquad (4)$$

where, X′ is a real number within the interval [a, b].

Consequently, OWOA employs a similar approach of opposition-based initiation, as seen in Figure 2 flowchart..

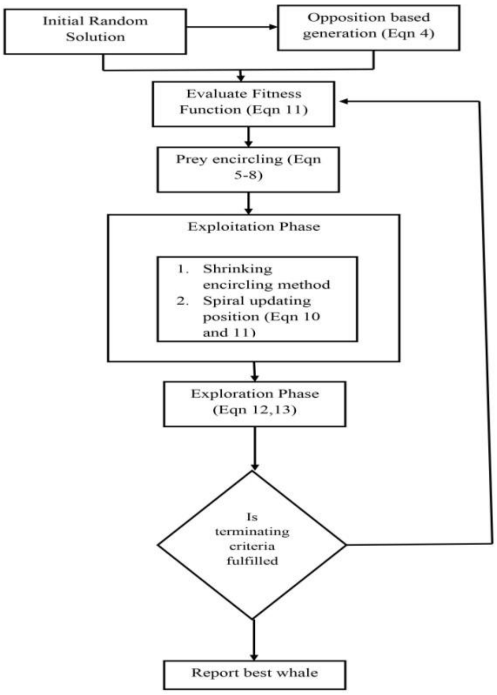

Initial Random Solution
Opposition based generation (Eqn 4)
Evaluate Fitness Function (Eqn 11)
Prey encircling (Eqn 5-8)
Exploitation Phase
1. Shrinking encircling method
2. Spiral updating position (Eqn 10 and 11)
Exploration Phase (Eqn 12,13)
Is terminating criteria fulfilled
Report best whale

Fig 2. Flow Chart of Proposed Work

Objective Function

The objective function f(x) assesses the performance of the voice recognition system based on a certain parameter configuration x. It may include measurements such as the accuracy of recognition, the Word Error Rate (WER), or other pertinent indicators.

Levy Flight

The Levy flight algorithm is used to induce stochasticity and promote thorough exploration inside the search domain. The whale locations are updated using step lengths generated by a Levy distribution.

Amplitude Coefficient A

The amplitude coefficient determines the magnitude of the step taken in the search space. It affects the extent of modifications made to the locations of whales throughout each cycle.

Opposition Strategy

Each whale's opposition positions are computed by the opposition technique and used to explore other search space areas, therefore improving search efficiency.

Convergence Criteria

The convergence criteria are used to define the point at which the optimization process should be terminated. This may be determined by either reaching the maximum number of iterations or attaining the required degree of solution quality.

Prey Encircling

Humpback whales possess the capacity to detect the whereabouts of their prey and encircle them. Since the ideal framework's exact location in the search space is unknown, the WOA method assumes that the best possible configuration right now is the target prey or very near to it. The search agent with the highest level of superiority is selected, and the other search agents make efforts to adjust their positions to match that of the top search agent. The following equations are relevant to this behavior:

$$E = |P.X^{*}(t) - X(t) \qquad (5)$$

$$Y(t+1) + Y^{*}(t) - Q.E \qquad (6)$$

In this context, t represents the current iteration, E represents the best solution achieved up to this point, Q and P are coefficient vectors, Y* represents the location of the best search agent, and Y represents the position vector. The vectors Q and P are computed.

$$B = 2c.f - c \qquad (7)$$

$$\text{D}= 2.f \qquad (8)$$

Algorithm 1: Opposition Whale Optimization Algorithm for Speech Recognition

Initialization

1. Initialize Population

Create a starting set of possible solutions that reflect different parameter settings for the voice recognition system.

2. Define Objective Function

Construct an objective function, denoted as f(x), which quantifies the performance of each potential solution x based on its accuracy in voice recognition.

3. Set parameter

Specify parameters such as the size of the population (N), the maximum number of iterations (max_iter), and any additional parameters unique to the method.

4. Randomization

Stochastically assign the beginning location of each whale inside the search area.

Main Loop

5. Iterative Optimization

Repeat for t=1 to max_itreation

For each whale i in the population:

- Evaluate the objective function f(x_i) to determine the fitness of the current solution.
- Update the position of the whale based on its current position and the positions of other whales using the following equation $x_i(t+1) = x_i(t) + A.rand\ ().Levy()$ (9)

Where

- A is the amplitude coefficient.
- rand() is a random number between 0 and 1.
- Levy() represents Levy flight for exploration.

If the updated position $x_i(t+1)$ is outside the search space, bring it back within the bounds.

6. Opposition Strategy

Once the locations have been updated, find out where each whale is in relation to the others by using:

$$x_{opp} = x_{max} + x_{min} - x_i \qquad (10)$$

- x_{max} is the maximum bound of the search space.
- x_{min} min$x$min is the minimum bound of the search space.

7. Evaluation

Assess the suitability of the opposing positions based on their fitness, denoted as f(xopp).

8. Update Best Solution

Revise the most optimal solution discovered so far by considering the fitness values of both the original and opposing positions.

9. Convergence check

Evaluate the convergence criteria. Terminate the loop if a certain condition is fulfilled; else, go to the next iteration.

Termination

Return Best Solution: Provide the optimal solution achieved either by reaching the maximum number of iterations or by meeting the convergence requirements.

3.3 Neural-Based Opposition Whale Optimization Algorithm (NOWOA)

Several techniques have been used and executed in previous instances to convert individual word voice into written text. The proposed SDRN incorporates handcrafted attributes. The AMS technique is used to extract the features from the input speech stream. The input qualities that were obtained are then used in a Deep Neural Network (DNN) to train and assess the model. The OWOA optimization approach is used to enhance the hyperparameter learning of deep neural network neurons. In this study, the approach shown in Figure 3 is called a neural-based opposition whale optimization algorithm (NOWOA). The latest optimization technique suggested is the OWOA algorithm.

Whales are often regarded as highly intelligent creatures. WOA is a recently developed optimization approach that is based on the behavior of humpback whales. This algorithm exhibits a superior degree of accuracy, and the opposing algorithm is included to further enhance its performance. In this study, the Convolutional Neural Network (CNN) is provided with 375 characteristics extracted from the input voice sounds. This research utilizes both standard and real-time databases.

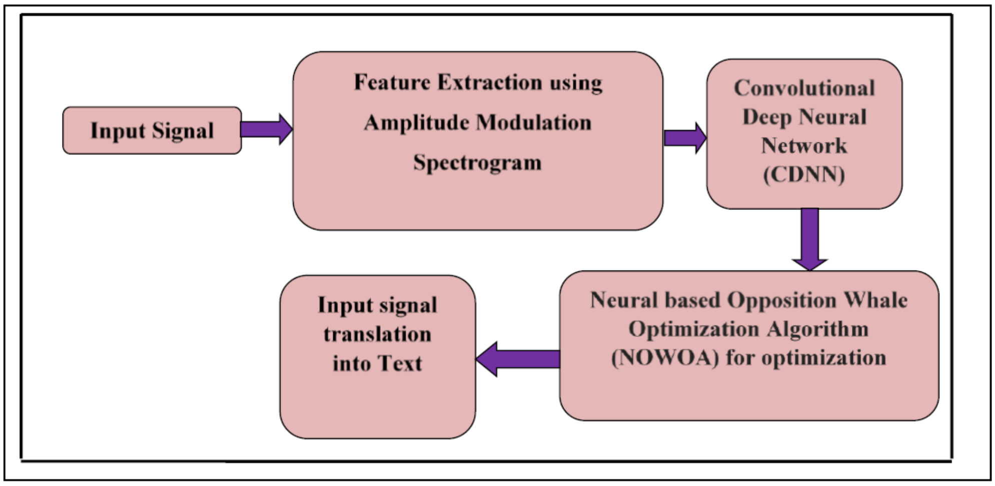

Input Signal
Feature Extraction using Amplitude Modulation Spectrogram
Convolutional Deep Neural Network (CDNN)
Neural based Opposition Whale Optimization Algorithm (NOWOA) for optimization
Input signal translation into Text

Fig 3. Basic work flow of Proposed work

3.3.1 Optimization of Hidden Layers and Neurons

The Whale Optimization Algorithm (WOA), Artificial Bee Colony (ABC), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Opposition-Based Whale Optimization Algorithm (OWOA) are among the strategies used to improve Deep Neural Network (DNN) layers and their neurons. In this technique, the WOA algorithm is adapted to OWOA to achieve optimum detection performance.

3.3.2 Original Solution Creation

The HL is indexed from 1 to 5, and the relevant neurons are indexed from 1 to 30 to get the first outcome.

3.3.3 Fitness Computation

The fitness function ff_i is given as:

$$ff_i = \frac{final\ correct\ Prediction\ data}{Total\ data} \quad (11)$$

The neural network topology is modified and the outcome is predicted using this fitness function.

4. Result and Discussion

4.1 Dataset

The TIMIT [37] corpus of reading discourse aims to provide acoustic-phonetic studies, event sequencing, and evaluation of automated discourse recognition systems using discourse information. TIMIT comprises recordings of six hundred and thirty individuals representing eight major dialects of American English, each reading 10 phonetically diverse sets of phrases. The TIMIT corpus includes orthographic, phonetic, and word records that are modified for time. It also includes a 16-bit, 16 kHz speech waveform file for each word. The TIMIT corpus recordings have been manually verified. After taking phonetic and provincial inclusion into account, we establish test and preparation subsets. Only plain PC accessible data and written documentation are included.Out of the total 6300 speech signals, 70% are used for training purposes and the remaining 30% are used for testing. A total of sixty distinct discourse signals are captured in real-time under various environmental settings for the purpose of validation. Out of these, seventy percent are used for training and the remaining thirty percent are used for testing. A total of 110 discourse signals were examined to identify consistent and uninterrupted discourse signals, with 70% of them allocated for training and 30% for testing. A total of 375 characteristics are extracted from these speech data. AMS is used to extract characteristics from the information included in the voice corpus initially. The input comprises a mixture of pristine and distorted signals that have undergone standardization, quantization, and windowing.

The signals captured were separated into different time-frequency (TF) units by using bandpass filters to modify the signals within a certain frequency range. The signals are divided into 25 TF units, each of which is assigned to channel Ci with i=1, 2, 3……, 25. There are a total of 25 groups of channels that have been evaluated, with each channel's signal frequency falling within the range of its own group.

Let (λ, υ) represent the component, where υ and λ correspond to different time frames and various sub-bands, respectively. Given the minimum changes in TF spaces, the capacity Δfti are viewed as the provisions that will be deleted. These provisions are represented by ti and B, which stand for time and channel data transmission, respectively.

$$\Delta f_{ti}(\lambda, \upsilon) = \mathrm{f_r}(\lambda, \upsilon) - \mathrm{f_r}(\lambda, \upsilon - 1) \tag{3.8}$$

$$where\ \upsilon = 2, \dots \dots .\,\mathrm{t_i}$$

The frequency delta function Δfd is given as:

$$\Delta \mathrm{fd}(\lambda, \upsilon) = \mathrm{f_r}(\lambda, \upsilon) - \mathrm{f_r}(\lambda - 1, \upsilon) \tag{3.9}$$

$$where\ \lambda = 2, \dots \dots .\,\mathrm{t_i}$$

The cumulative FV $f\,(\lambda, \upsilon)$ can be defined as:

$$f(\lambda, \upsilon) = [\mathrm{f_r}(\lambda, \upsilon), \Delta \mathrm{f_{ti}}(\lambda, \upsilon), \Delta \mathrm{f_d}(\lambda, \upsilon) \tag{3.10}$$

4.2 Quality Parameter

The classification graph is shown by using the following performance assessment criteria for these five approaches:

Specificity

Specificity, commonly referred to as the True Negative Rate, is a measure used in binary classification tasks to assess a model's capability to accurately recognize negative cases. The metric quantifies the accuracy with which the model properly classifies real negative occurrences, i.e., the actual negative cases.

$$Specificity = \frac{True\ Negaives}{True\ Negative + False\ Positive} \tag{3.11}$$

Accuracy

Accuracy is a metric that calculates the ratio of properly identified examples to the total number of occurrences in a dataset.

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ Predictions} \tag{3.12}$$

Sensitivity

Sensitivity, sometimes referred to as the True Positive Rate or Recall, is a measure used in binary classification tasks to assess a model's capacity to accurately detect positive cases. The metric quantifies the accuracy of the model in properly classifying real positive cases, also known as true positives.

$$Sensitivity = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ Predictions} \tag{3.13}$$

Negative predictive value (NPV)

The Negative Predictive Value (NPV) is a quantitative measure used in binary classification tasks to assess the model's accuracy in properly predicting negative occurrences where the projected result is negative. NPV quantifies the accuracy of the model in correctly identifying negative cases out of all the instances predicted as negative.

$$NPV = \frac{True\ Negative}{True\ Negative + False\ Negatives} \tag{3.14}$$

False-negative rate (FNR)

The False Negative Rate (FNR), also known as the Miss Rate, is a metric used in binary classification tasks to quantify the proportion of actual positive examples that are incorrectly classified as negative by the model.

$$FNR = \frac{False\ Negative}{False\ Negative + True\ Positives} \tag{3.15}$$

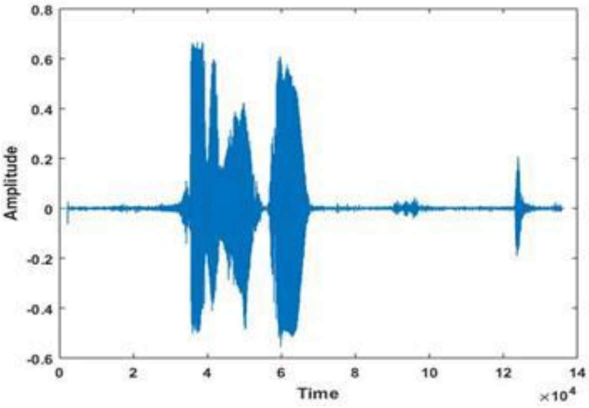
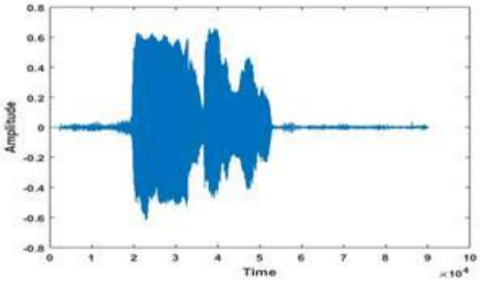
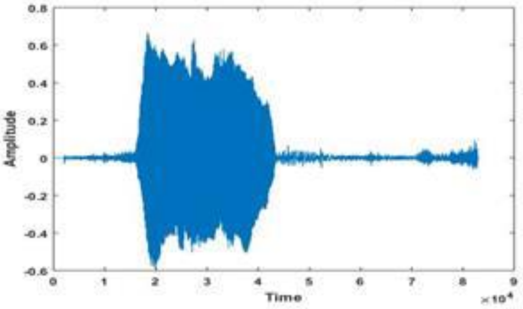
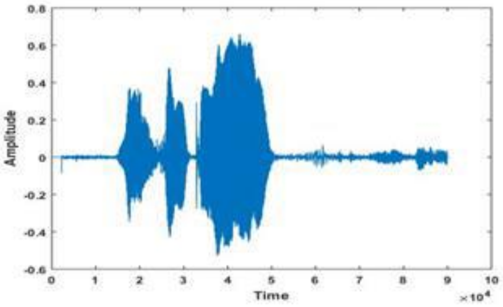
False-positive rate (FPR)

The False Positive Rate (FPR), sometimes known as the False Alarm Rate, is a metric used in binary classification tasks to quantify the proportion of actual negative instances (or correctly recognized negatives) that are incorrectly classified as positive by the model.

$$FNR = \frac{False\ Negative}{False\ Negative + True\ Positives} \tag{3.16}$$

The voice recognition technique involves analyzing unique speech signals from various individuals, such as phrases like 'Have a great day', 'How are you', 'How old are you', 'Welcome to all', and 'What is your name'. The artificial neural network (ANN) classification approach is used to make predictions on the text. The LM algorithm is executed for training purposes, with the objective of optimizing the HL and neuron. To do this, three optimization algorithms are employed: WOA, OWOA, and OABC. These three algorithms are compared using a classification graph that displays their respective performances on metrics like accuracy, sensitivity, specificity, NPV, FPR, FNR, and FDR. Table 1 displays a range of input speech and amplitude signals.

Table 1. sample Input speech from dataset

Input Speech	Amplitude Signal
Have a Nice Day	
How Old Are You	
How Are You	
Welcome to all	

The optimization process utilizes the database to forecast the most suitable CNN design that will result in the highest recognition accuracy. Table 2 presents a concise overview of the relevant information on the expected architectures of CNNs. The text provides a visual representation of the quantity of neurons in each hidden layer of the CNN architecture for different optimization strategies.

Table 2. CNN Hidden Layer design for different Optimization

S.No	Optimization Techniques	No of neurons in HL-1	No of neurons in HL-2	No of neurons in HL-3
1	WOA	16	23	24
2	ABC	22	20	-
3	PSO	9	31	-

4	GA	6	19	-
5	NOWOA	23	19	-

The tables demonstrate that the proposed NOWOA technique exhibits a lower HL (Hamming Loss) and that the neurons have been fine-tuned to achieve optimal identification accuracy. The outcome unequivocally demonstrates that the use of the suggested NOWOA enables the modification of the architecture of DNN to achieve optimal detection accuracy. This can be achieved by correctly training and testing the DNN with any kind of dataset, including real-time data. Table 3 displays the training durations used in the algorithms.

Table 3. Training time for each optimization

S.no	Optimization Algorithm implementation	Time in sec
1	WOA	1513.38
2	ABC	1462.8
3	GA	1403.7
4	PSO	16.08.8
5	NOWOA	13.04.2

Table 3 demonstrates that NOWOA exhibits a shorter training time in comparison to the other algorithms. Consequently, it optimizes computational efficiency, minimizes costs, and enhances the speed of identification.

Figure 4 demonstrates the real-time examination of NOWOA's performance on individual signals. It surpasses other often used methods. The proposed approach demonstrates outstanding performance across all evaluated parameters, thanks to the implementation of a neural-based opposition strategy. In terms of precision, NOWOA achieved a score of 97.1 percent, surpassing WOA by 1.7 percent, ABC by 2.7 percent, PSO by 6.4 percent, and the least accurate GA by 12.9 percent. In contrast to earlier optimization approaches, NOWOA also demonstrates outstanding performance on other conventional metrics.

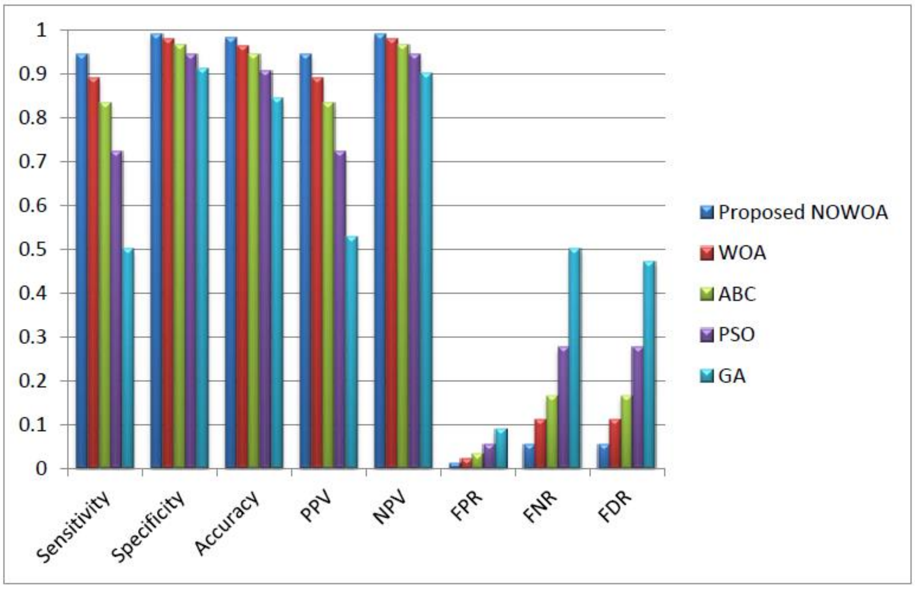


Fig 4. Comparison of various parameters among different optimization

Figures 5 provide a comprehensive comparison of the key features of the proposed NOWOA approach with the WOA and NOABC algorithms for continuous voice signals.

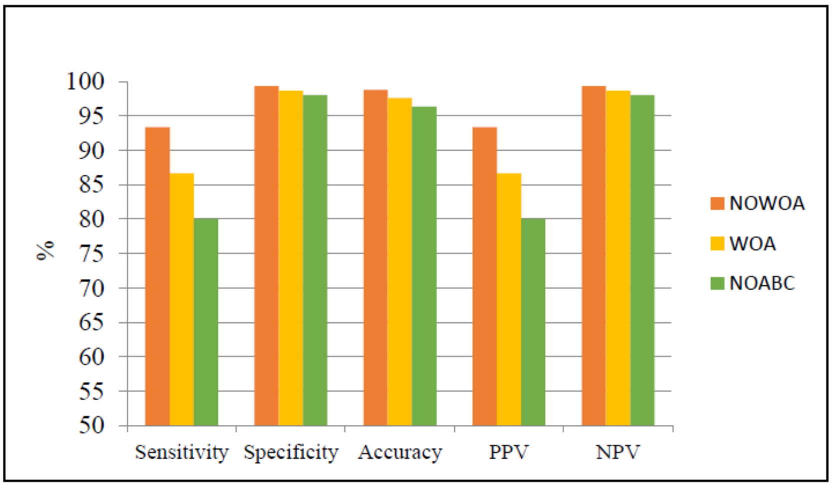


Fig 6: Performance Comparison

Figure 7 presents the performance assessment of the proposed NOWOA in comparison to several current approaches. The analysis of recognition accuracies reveals that the maximum accuracy is achieved by NOWOA, reaching 95.4%. The recognition accuracy of NOABC and GA-based fuzzy logic neural networks [9] is 91.36% and 83%, respectively. The CSO-based Artificial Neural Network (ANN) achieves a recognition accuracy of 89.65%, while the Neural Network (NN) based system using Hidden Markov Model (HMM) classifier and Probabilistic Input-Output (PIO) achieves an accuracy of 80% [10]. The findings demonstrate that the suggested approach surpasses the performance of other current processes.

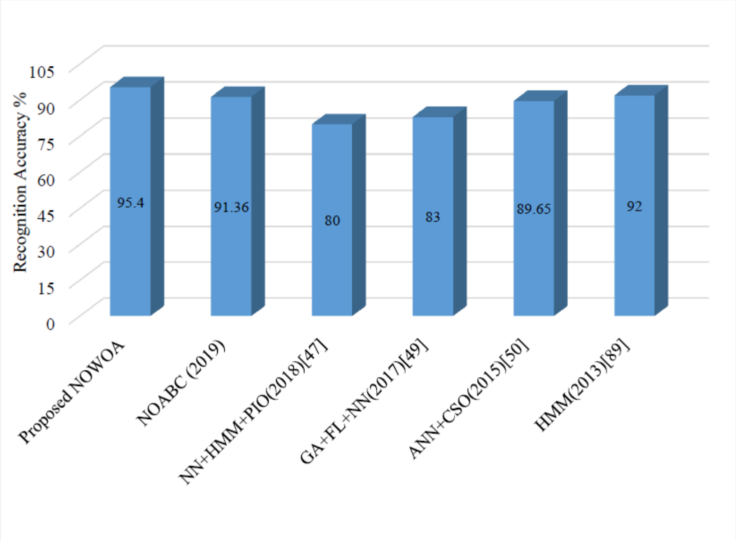


Fig 7. Performance comparison of proposed Vs Existing method

5. Conclusion

The suggested SDRN is unique because it enhances the architecture of the NN (hidden layers and their neurons) by including NOWOA. This study identified individual words and uninterrupted voice signals and transcribed them into written text. AMS is used to extract characteristics from diverse communication streams. The study findings indicate that the accuracy value of NOWOA is 97.1. The result demonstrates that NOWOA effectively identifies the optimal set of characteristics while retaining high classification accuracy.

References

[1] Liu, Y., Karanasou, P., Hain, T.,“ An investigation into speaker informed DNN front - end for VCSR ,” In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 4300 - 4304, April., 2015.

[2] Sainath, T.N., Kingsbury, B., Mohamed, A.R., Dahl, G.E., Saon, G., Soltau, H., Beran, T., Aravkin, A.Y., Ramabhadran, B., “ Improvements to deep convolutional neural *networks for LVCSR*,” In 2013 IEEE workshop on automatic speech recognition and understanding, pp. 315-320, December 2013.

[3] Sakshi Dua, Sethuraman Sambath Kumar, Yasser Albagory, Rajakumar Ramalingam , Ankur Dumka, Rajesh Singh, Mamoon Rashid , Anita Gehlot, Sultan S. Alshamrani and Ahmed Saeed AlGhamdi, “Developing a Speech Recognition System

forecognizing Tonal Speech Signals Using a Convolutional Neural Network", Appl. Sci. **2022**, 12, 6223. https://doi.org/ 10.3390/app12126223.

[4] Jagannath Dayal Pradhan, L. V. Narasimha Prasad, Tusar Kanti Dash, Manisha Guduri, and Ganapati Panda, "Cascaded PFLANN Model for Intelligent Health Informatics in Detection of Respiratory Diseases from Speech Using Bio-inspired Computation", Journal of Artificial Intelligence and Technology, 2024, 4, 124-131.

[5] Tusar Kanti Dash, Soumya Mishra, Ganapati Panda, Suresh Chandra Satapathy , "Detection of COVID-19 from speech signal using bio-inspired based cepstral features", www.elsevier.com/locate/patcog, Available online 24 April 2021.

[6] Rajni Sobti, Kalpna Guleria, Virender Kadyan, " Automatic Speech Recognition System for Low Resource Punjabi Language using Deep Neural Network-Hidden Markov Model (DNN-HMM)", International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING, 2024.

[7] Li, X.; Zhou, Z. Speech command recognition with convolutional neural network. In CS229 Stanford Education; 2017. Available online: http://cs229.stanford.edu/proj2017/final-reports/5244201.pdf

[8] Taniya Hasija, Virender Kadyan and Kalpna Guleria, "Out Domain Data Augmentation on Punjabi Children Speech Recognition using Tacotron", Journal of Physics: Conference Series, 1950 (2021) 012044, IOP Publishing, doi:10.1088/1742-6596/1950/1/012044.

[9] Waris A., Aggarwal R.K., *"Optimization of deep neural network for automatic speech recognition,"* In International Conference on Inventive Research in Computing Applications (ICIRCA), India, pp. 524-527, 2018.

[10] Jayasankar T., Vinothkumar K., Vijayaselvi, A. *"Automatic gender identification in speech recognition by genetic algorithm,"* Applied Mathematics & Information Sciences, vol.11, no.3, pp. 907-913, 2017.