A Novel Optimized Deep Learner with Lightweighted Encryption-Based Speaker Verification

Sujiya Rathinaraja (suji.sreedharan@gmail.com)
Bharathiar University School of Computer Science and Engineering
Chandra Eswaran
Bharathiar University School of Computer Science and Engineering

Research Article

Keywords: Speaker authentication, MGWOVSW-CAES-GMM, Elephant herding optimization, Convergence rate, DNN

Posted Date: June 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-486319/v1

A Novel Optimized Deep Learner with Lightweighted Encryption-based Speaker Verification

Sujiya Sreedharan, PhD Research Scholar, Department of Computer Science, Coimbatore, India, suji.sreedharan@gmail.com
Chandra Eswaran, Professor & Head, Department of Computer Science, Coimbatore, India, crcspeech@gmail.com

ABSTRACT—

In the modern digitalized world, Speaker verification (SV) system is essential for authorizing the client's credentials. To design an effective SV system, MGWOVSW-CAES-GMM system has been proposed. In this system, the Modified Grey Wolf Optimization (MGWO) technique was employed to optimize the variable sliding window size, FMPM features and training variables. The optimized features were watermarked and encrypted using a Chaotic-based Advanced Encryption Standard (CAES). Once the encryption process was completed, the encrypted features were forwarded to the recipient who executes the decryption and de-watermarking processes. At last, the decrypted features were classified using Gaussian Mixture Model (GMM) classifier. Conversely, MGWO has poor convergence rate and ineffective searching results. Hence, this article proposes an EEHOVSW-CAES-GMM system in which Enhanced Elephant Herding Optimization (EEHO) algorithm is applied instead of MGWO. On the contrary, the computational complexity of GMM classifier is high and its efficiency is less while increasing the number of features. For this reason, a Deep Neural Network (DNN) classifier is employed instead of GMM for recognizing the decrypted features and authorize the speaker's identity. Besides, the parameters utilized in DNN topology are optimized using two different systems such as MGWOVSW-CAES-DNN and EEHOVSW-CAES-DNN for reducing the computational complexity and increasing the classification accuracy effectively when using more number of features. By using these classifiers, the speaker's identity is verified and the attacks during the transmission are prevented with the highest security level.

Keywords—Speaker authentication, MGWOVSW-CAES-GMM, Elephant herding optimization, Convergence rate, DNN

I. INTRODUCTION

Speaker authentication is usually the method of authenticating a speaker's potential functionalities depending on the voice utterance properties received from several audio devices. The primary concept of SV comes within the well-known component of speech authentication and it can partition into text-dependent and text-independent. In particular, it extracts Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) features in the short-term frequency of voice utterances [1-2]. As well, Frequency Domain Linear Prediction (FDLP) coefficients are used to transform the speech utterances to the spectral field by the Discrete Cosine Transform (DCT). Additionally, it extracts the Mean Hilbert Envelope Coefficients (MHEC) and Power-Normalized Cepstral Coefficients (PNCC) [3]. The variability among these features is correctly estimated through the covariance of the pre-computed cepstral features which varies according to the vocalized words, transmission medium and locale noise. In classical methods, a voice uncertainty is computed by the GMM's covariance matrices [4]. The use of GMM is centered on the Gaussian basic aspects which reflect a small number of familiar spectral characteristics based on voices like Gaussian potentials to predict variable frequencies.

During the past years, the absolute variations in feature space were primarily related to the spoken words. However, the exploration of such variations may support for SV as every speaker has their articulatory traits. By using the global covariance matrix, the neighbourhood covariance with various identified transmission fragments of many amplifiers was analysed [5]. As a result, the spectral features were extracted via the pre-computed range of sliding window. On the other side, few significant features were lost during extraction due to the pre-computed window dimension which is not appropriate for all kinds of speech corpus. As well, the feature extraction was complex because of a volatile nature of speech and distorted transmission channels. To avoid these problems, an OVSW-SV system was suggested that extracts the feature called FMPM [6] using FDLP, MHEC, PNCC and MFCC with VSW. Also, MGWO was applied for

optimizing the SW size, choosing the best characteristics and training variables in an adaptive way.

Once the best features were chosen, the GMM-Universal Background Model (UBM) classifier was applied for recognition task. In contrast, cryptographic mechanisms must be integrated to guarantee the confidentiality against many communication threats. As a result, an Improved Encryption (IE)-OVSW-SV system [7] was proposed that incorporates an enhanced AES encryption scheme for authenticating the transmitted speech signal features from sender to the receiver. In this system, the best features were digitalized, framed and reshaped. Then, all features were watermarked with the help of image and its index was inserted into the speech utterances as the Least Significant Bits (LSB) in the primary 8 bytes. Afterwards, the resultant features were cipher texted by an enhanced AES scheme wherein an enhanced RC4 codeword creator was used for creating the stream cipher codeword. This codeword was then applied for encrypting the features which are forwarded to the recipient who executes de-watermarking and deciphering schemes for acquiring the underlying voice utterance properties. Further, these features were given to the GMM-UBM classifier for verifying the legitimate identities. Conversely, the enhanced AES was symmetric cryptographic scheme which has high computational cost and complexity.

To combat these challenges, a MGWOVSW-CAES-GMM system has been designed that uses chebyshev chaotic map and enhanced quadratic map to encrypt the voice utterance features with the highest security level [8]. In this system, the selected best speech signal features were encrypted using shortened AES algorithm that eliminates few rounds of permutation and also employs an enhanced quadratic chaotic map enciphering for increasing the confidentiality level. This algorithm consists of 2 major tasks such as permutation and diffusion. Permutation can relocate the enciphered features with no modification of the underlying ranges while diffusion can encipher the features via modifying the underlying ranges using Chebyshev chaotic map [9]. Moreover, the chaotic series of enciphered features given by the chaotic logistics map with chebyshev polynomial map was sorted to get the permutated features. After that, these were enciphered by the codeword associated with the given input via enhanced

quadratic chaotic maps because it needs just single cycle of diffusion; yet, the traditional quadratic chaotic maps needs at least 3 cycles of diffusion.

Once the features were encrypted, these were sent to the receiver where the inverse processes were performed to get the original features. Further, GMM-UBM classification was employed to verify the customer's credentials. On the other hand, the convergence and searching capabilities of MGWO were not effective. Thus, in this article, an EEHOVSW-CAES-GMM system is proposed in which EEHO algorithm instead of MGWO to achieve fast convergence. This EEHO enhances the convergence rate through modifying the group and partition operations. Additionally, a fixed variable is considered to analyse the biased convergence of EEHO. However, GMM-UBM classifier has high computational complexity and does not operate well while increasing the number of features. Also, the user should set the amount of mixture models that the algorithm can fit to the training dataset. When the number of features is increased, the user may not capable to decide the number of mixture models to achieve the best classification results.

As a result, DNN classifier is proposed instead of GMM for verifying the speaker's identity. Moreover, the DNN parameters are fine-tuned via MGWO and EEHO algorithms that can reduce the computational complexity and achieve the best classification results while increasing the number of features or instances. These two different proposed systems are named as MGWOVSW-CAES-DNN and EEHOVSW-CAES-DNN. Thus, these proposed systems can enhance the classification of encrypted and decrypted speech signals for authenticating the valid speakers.

The remaining sections are involved the following: Section II discusses the previous works associated with the SV systems using different algorithms. Section III explains the methodologies of proposed systems and Section IV shows their efficiencies. Section V concludes the entire work.

II. LITERATURE SURVEY

Novoselov et al. [10] proposed a two-deep non-linear Probabilistic Linear Discriminant Analysis (PLDA) for i-vector SV. In this method, direct utilization of class labels was

employed for solving the two PLDA processes. Based on this method, within-class variability was reduced and between-label variability was maximized. However, this strategy did not suit all kinds of audio corpus. Soleymanpour & Marvi [11] proposed text-independent SV depending on the decision of the maximum identical features. The MFCC of each window over given voice utterance has been used as features and the features having the highest similarity were acquired via the clustering technique. After that, the selected features have been recognized by the Artificial Neural Network (ANN). Nevertheless, the accuracy of this system was not effective.

Thullier et al. [12] proposed a text independent SV technique for cellular phones through amplifiers. This technique has the aim of improving the SV system via three fundamental tasks. At first, a group of user's voice characteristics was extracted to generate the training set. Then, it was classified via a naive Bayes classification. Further, a final decision was estimated for verifying the speaker. However, the precision has not yet improved effectively.

Dey et al. [13] proposed a template mapping for text-dependent SV with the goal of achieving phone sequence information using Dynamic Time Warping (DTW) with audio-useful properties. These were acquired from i-vector frameworks mined over short audio portions. Additionally, PLDA was proposed for projecting the online i-vectors onto the user-discriminative subspace. Conversely, the linguistic data in the voice utterance was not considered.

Chowdhury et al. [14] introduced an attention layer as a soft technique for emphasizing the most relevant components of the input sequence. The main aim of this method was using a distributed-variable non-linear ranking measure, using a split- layer attention link to the final layer of the Long Short-Term Memory (LSTM) and applying the SW max-pooling on the attention weights. However, the computational complexity of this method was high.

Snyder et al. [15] proposed the data augmentation for increasing the DNN embeddings efficiency for SV. This method was proposed as a low-cost method that consists of added noise and reverberation for multiplying the number of training data and

increasing the robustness. Conversely, the performance was still not effective in terms of detection accuracy. Dey et al. [16] proposed a novel distance factor for authenticating the voice utterances. First, a dynamic size audio fragment has been synchronized to the pre-determined size features through calculating the mean secret interpretations in the DNN. The gap among voice signals has been determined using l_2-norm. Additionally, the audio metadata has been merged based on the gap value along dialect components between registration and validate set. Further, the whole structure has been fine-tuned by means of triplet-loss to calculate the SV grades. However, the performance was worse than GMM-UBM classifier.

Meng et al. [17] proposed an adversarial SV method for learning the criterion-invariant deep embedding through adversarial multi-task learning. First, a user categorization model and a criterion detection model were modified together for reducing the user categorization error and concurrently min-maximizing the criterion error. Moreover, multi-factorial adversarial SV was proposed for concurrently suppressing many factors that form the condition variability. Further, a regression model was employed for restoring the constant criterion factor. Conversely, the Equal Error Rate (EER) of this method was high.

Gupta et al. [18] designed an enhanced two-level scheme in which the first-level recognizes the user's gender and the second-level verifies the user according to their gender. Once the speaker's gender was recognized, hunt space was minimized that explores in a group of voice utterances regarding the recognized gender. The MFCC and pitch were used for gender recognition using Support Vector Machine (SVM) while the MFCC, pitch and Relative Spectral-Perceptual Linear Predictive (RASTA-PLP) were used for SV using GMM classifiers. However, the threshold must be calculated to select appropriate characteristics.

Abualadas et al. [19] designed hybrid feature extraction methods for SV. Initially, the influences of proper selected characteristics at different stages of DWT were studied. Then, the merging of DWT and curvelet transform methods was proposed for extracting the features. Once all the features were extracted, Principal Component Analysis (PCA) was performed for lessening the number of features. Finally, Back-Propagation (BP)

neural network was utilized as the classifier for verifying the speaker. However, few irrelevant and noisy features were not removed effectively that degrades the accuracy.

Shon et al. [20] proposed a new loss function, namely VoiceID loss for training the speech enhancement framework that increases the robustness of SV. This VoiceID error was depending on the response from the SV framework for generating the percentage mask. It was then multiplied pointwise by means of the actual spectrogram for filtering redundant elements. Nonetheless, this framework does not consider the acoustic features i.e., MFCC that can degrade the robustness of the system.

Liu et al. [21] extracted the phonetic vectors from a modified automated SV framework and considered as secondary incomings into the x-vector structure. Then, hybrid multi-task training was applied to obtain the distributed data between user and linguistic characteristics. At last, c-vector framework was proposed that utilizes the linguistic variables during training. However, efficiency has been diminished because it cannot offer greater valuable data and attribute removal has been complicated whilst amount of layers in the linguistically discriminating system has declined.

Kakade & Salunke [22] proposed automated voice-user verification in real-world configurations. In this system, noise from the sampled voice signal was removed via pre-processing. Then, MFCC technique was used for extracting the features. During testing, all voice utterances were harmonized by the Vector Quantization (VQ) and DTW methods. Finally, the speaker and speech were verified through the results obtained from the mapped VQ and DTW, accordingly. However, it did not examine the verbal data of a given person and the computation efficacy has been decreased by the utilization of numerous features.

Arora &Vig [23] proposed effective SV for voice signal from cellular systems with the aim of extracting, characterizing and verifying the data about speaker identity. This system has four phases such as utterance splitting, feature extraction, feature choice and verification. First, an utterance splitting scheme was used for shortening the voice into many short-duration signals. Then, artifact elimination was applied and the Mel Advanced Hilbert-Huang Cepstral Coefficients (MAHCC) method was employed to

extract the characteristics in a considered voice. Additionally, the crow search algorithm was executed to choose the appropriate characteristics through ranking. At last, the Deep Hidden Markov Model (DHMM) was employed to verify the user. On the contrary, error probability and computation difficulty were high.

Sun & Chol [24] focused on the flexible series voice modelling method of fusing more contextual data of Recurrent Neural Network (RNN) with alteration capacity of Subspace GMM (SGMM) for automated SV. The main objective of this method was obtaining the prior features of a phrase with more contextual data and modeling each phrase by SGMM applied in voice modeling for automated SV. In addition, the variable merging method was applied to obtain the SGMM. However, the time taken to measure the likelihood was high. Ramteke et al. [25] segmented and extracted the vowel features for each desired and undesired voice. Further, the random forest, SVM and Deep Feed Forward Neural Network (DFFNN) were applied to authenticate the users with the help of extracted features. On the contrary, the computational complexity was high.

III. PROPOSED METHODOLOGY

In this section, the proposed SV systems using different kinds of classifiers are briefly explained. The core functionalities in the proposed SV systems shown in Figure 1 are the following:

- *Training Phase*:

1. Primarily, FMPM characteristics are mined depending on the adapting the VSW and the highly appropriate characteristics are elected by the EEHO scheme.
2. After, those elected characteristics are reformed into 4×4 matrix (A) of bytes.
3. The first eight bytes of features in each block are watermarked using the image which is XORed with each block [7].
4. After that, CAES algorithm [8] is applied for encrypting the features and the enciphered features are broadcasted to the recipient with the maximum security level.

5. The recipient conducts the de-watermarking and deciphering tasks to get the actual characteristics.
6. Moreover, the GMM & DNN classifiers are applied to classify the decrypted features. Here, the classification variables are adjusted using the MGWO and EEHO schemes.

- *Testing Phase*:

1. The testing samples are given to the trained classifiers for evaluating the unknown speaker's identity and performance of the SV systems.

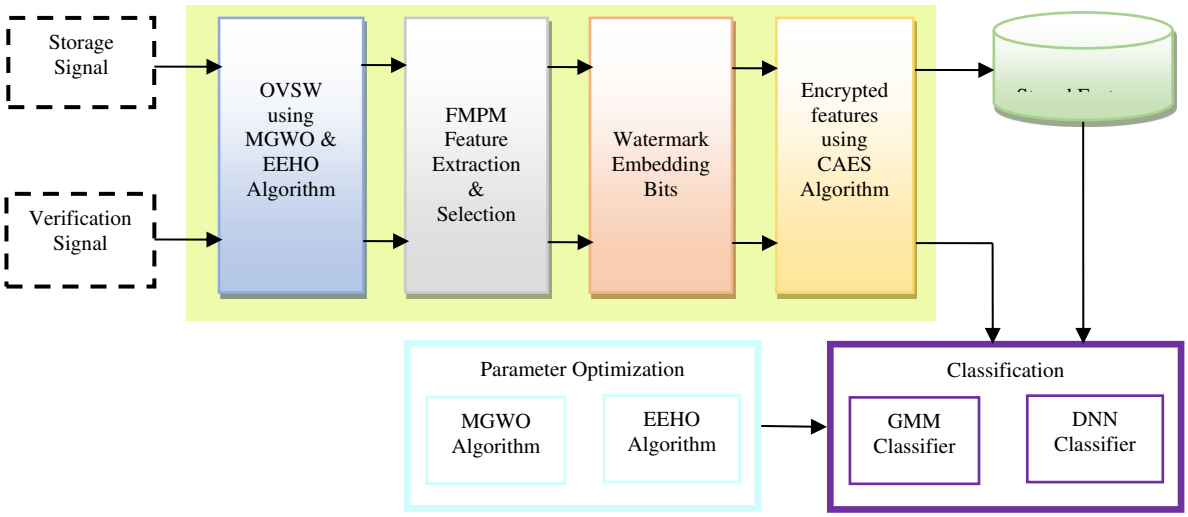

Figure 1. Schematic Representation of Proposed OVSW-SV Systems using Two Different Optimization Algorithms

3.1 EEHO Algorithm

It is applied to realize the high convergence rate by modifying the group and partition operations. In each population, an elephant having the highest strength in a group C_u is decided as the mother (m) at interval t.

$$m_u^t = \underset{e \in e_u}{\operatorname{argmax}} F(e) \qquad (1)$$

In Eq. (1), e_i denotes the group of elephants in C_u. All elephants v in group u covers a prior site $e_{u,v}^t$. Their fresh site $e_{u,v}^{t+1}$ is inclined by the group mother m_u^t as:

$$e_{u,v}^{t+1} = e_{u,v}^{t} + \alpha \times \left(m_u^t - e_{u,v}^{t}\right) + \beta \times \left(C_u^t - e_{u,v}^{t}\right) + \gamma \times r \qquad (2)$$

In Eq. (2), α, β and $\gamma \in [0,1]$ are the scaling variables which determine the influence of m_u^t on the elephant fresh site, the elephant's relation to travel towards the mid group and the elephant's relation to randomly walk, correspondingly, $r = (2 \times rand - 1)(e_{max} - e_{min})$ is an arbitrary vector acquired from the normal distribution, e_{max} and e_{min} are the upper and lower boundaries of the elephant's site and C_u^t is the group mid determined as:

$$C_u^t = \frac{1}{n_u} \times \sum_v x_{u,v}^t \qquad (3)$$

In Eq. (3), n_u denotes the amount of elephants in group u. The new location of the mother is not depending on its previous location. Also, for low ranges of β, the mother is rapidly travelled nearer to the source. For high ranges of β, the mother is rapidly travelled nearer to the group mid. For fixing the mother update function, $m^{t+1} = m^{t+1} + \beta(C^t - m^{t+1})$ must be utilized and the mother's fresh site is a linear mixture of its previous site. Here, $\alpha, \beta,$ and γ are used for controlling the convergence towards the group mid and an arbitrary walk in parallel. Since considering a single factor does not control the exploitation abilities.

The partition operation prepared by male elephants is modelled as:

$$e_{u,worst}^{t} = e_{min} + (e_{max} - e_{min}) \times rand \qquad (4)$$

In Eq. (4), $e_{u,worst}^{t}$ denotes the most unpleasant elephant in C_u. For the partition function, pseudorandom generator generates $rand$ i.e., arbitrary value between 0 and 1. For generating the normally shared arbitrary value between e_{min} and e_{max}, $rand$ is scaled and shifted. To achieve this, floor function is used i.e., $floor([e_{min}, e_{max}]) = [e_{min}, e_{max-1}]$; thus a series normal distribution between e_{min} and e_{max+1} is used for generating a discrete normal distribution between e_{min} and e_{max}.

Algorithm:

Initialization: Assign $t = 1$, the populace dimension pop_size and the maximum population Max_pop

$\boldsymbol{while}(t \leq Max\_pop)$

$\boldsymbol{for}(u = 1; u \leq nClan; u++)$

Sort group elephants depending on their strength;

$e_{u,best}^{t} = First\ elephant$;

$e_{u,worst}^{t} = Final\ elephant$

Utilize group modification operation;

Utilize partition operation;

Determine populace according to the fresh sites;

$t = t + 1$;

$\boldsymbol{end\ for}$

$\boldsymbol{end\ while}$

3.2 DNN Classifier with Optimization

The categorization efficiency is improved by a MGWO and EEHO which adjusts the variables used in the DNN with the help of MGWO and EEHO. Such algorithms are employed in the search for DNN's optimal topology which facilitates the capacity for modeling complicated nonlinear operations when automatically resolving the adaptation dilemma. Typically, DNNs illustrated in Figure 2 encompasses an input unit, a hidden unit and an output unit. Each unit has a group of neurons which are circulated over more stacked units and every unit is completely linked to the consecutive units.

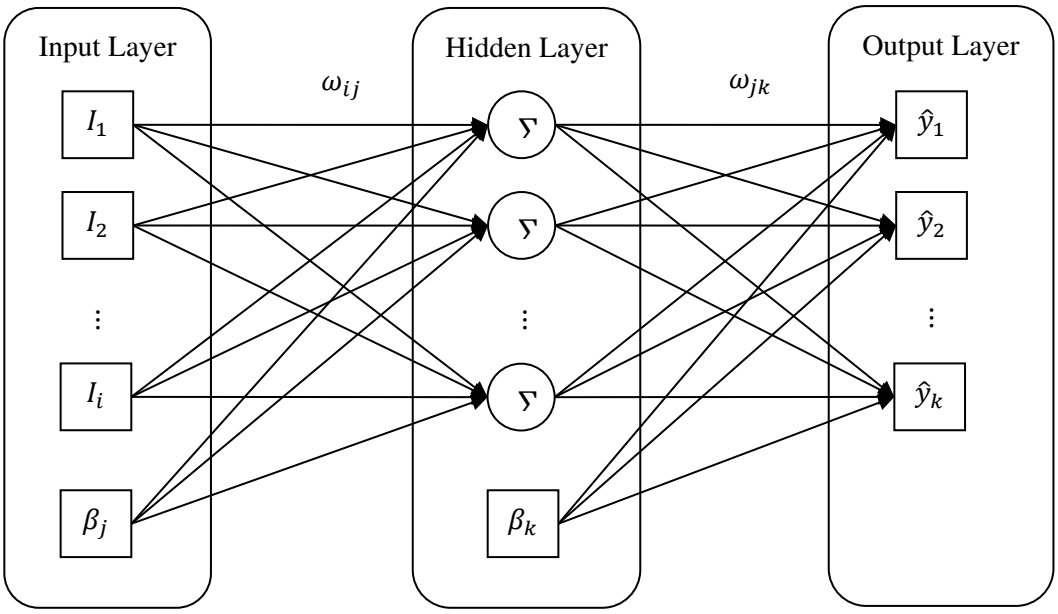

Figure 2. Schematic Structure of DNN Classifier

In DNNs, the neurons are linked in a single direction and the connections are specified by weights assigned between -1 and 1. The outcome of every neuron is computed using the following steps:

- Initially, the weighted summation of the input is computed as:

$$S_j = \sum_{i=1}^{n} \omega_{ij} I_i + b_j \qquad (5)$$

  In Eq. (5), I_i is the incoming features i, ω_{ij} is the weight between I_i and the hidden neuron j and also b_j is the bias.

- Then, an activation factor is applied for triggering the neuron's outcome depending on the summation operation. By using the sigmoid activation factor, the outcome of hidden neuron j is computed as:

$$f_j(x) = \frac{1}{1+e^{-S_j}} \qquad (6)$$

- Once the outcome of all hidden neurons is computed, the resultant outcome is computed as:

$$\hat{y}_k = \sum_{i=1}^{m} \omega_{kj} f_i + b_k \qquad (7)$$

Thus, the final output is used for classifying the decrypted features instead of GMM classifier for SV. However, it did not examine the optimized weights, bias, and the amount of hidden units whilst raising the overall network's depth. Thus, this is done by the MGWO [6] based on the hunting behaviour of grey wolves. On the other side, MGWO has slow convergence that affects the classification accuracy. So, DNN variables are successfully fine-tuned by the EEHO which also lessens the training loss.

Algorithm:

Input: Decrypted features

Output: Trained DNN

Initialize all weights and biases in the network;

$\boldsymbol{while}(termination\ condition\ is\ not\ satisfied)$

$\boldsymbol{for}(each\ input\ neuron\ i)$

$S_j = \sum_{i=1}^{n} \omega_{ij} I_i + b_j;$

$\boldsymbol{for}(each\ hidden\ layer\ neurons\ j)$

$f_j(x) = \frac{1}{1+e^{-S_j}};$

$\hat{y}_k = \sum_{i=1}^{m} \omega_{kj} f_i + b_k;$

$\boldsymbol{for}(each\ neuron\ k\ in\ output\ layer)$

$E_k = \frac{1}{2}\left(t_k^p - \hat{y}_k^p\right)^2$

//t_k^p is the desired target outcome for the p-th feature,$\hat{y}_k^p$ indicates the real outcome for the p-th feature and E_k indicates the error value.

$\boldsymbol{end\ for}$

$\boldsymbol{end\ for}$

$\boldsymbol{end\ for}$

if $(E_k \; is \; maximum)$

Initialize the number of hidden layers;

Initialize set of weights ω, biases β;

Implement MGWO algorithm or EEHO algorithm;

Optimize the number of hidden layers, weight and bias values;

end if

Train the DNN using the optimized parameters;

Return trained DNN;

end while

After training the DNN using optimized parameters, the testing samples are given to this classifier for verifying the speaker's identity. From this classification, it is noticed that the EEHOVSW-CAES-DNN system can minimize the classification error and computational complexity while increasing the number of speech signal features. As well, the fastest convergence is achieved that reduces the training time efficiently.

IV. RESULT AND DISCUSSION

This part presents the efficiency of EEHOVSW-CAES-GMM, MGWOVSW-CAES-DNN and EEHOVSW-CAES-DNN systems in terms of different evaluation metrics. Also, the efficiency is compared with the existing MGWOVSW-CAES-GMM classifier. The analysis is carried out with the help of Switchboard and TIMIT [6] corpuses. In this experiment, 180 speech utterances are taken into the consideration where 100 samples are recorded for training and 80 samples are recorded for testing. Among 100 training samples, 67 are male and 33 are female voices. Similarly, 50 are male and 30 are female voices in the testing samples.

Table 1 shows the performance results of evaluation metrics for classifying the speech signal features using MGWOVSW-CAES-GMM, EEHOVSW-CAES-GMM, MGWOVSW-CAES-DNN and EEHOVSW-CAES-DNN.

Table 1 Comparison of Classification Performance

Performance Metrics	MGWOVSW-CAES-GMM	EEHOVSW-CAES-GMM	MGWOVSW-CAES-DNN	EEHOVSW-CAES-DNN
Precision	0.8724	0.8832	0.8905	0.9018
Recall	0.9238	0.9314	0.9422	0.9576
F-measure	0.8981	0.9073	0.9164	0.9297
Accuracy (%)	97.5	98.1	98.6	99.3

4.1 Precision

It is the ratio that measures the classified features that are appropriateat True Positive (TP) values.

$$Precision = \frac{TP}{TP + False\ Positive\ (FP)}$$

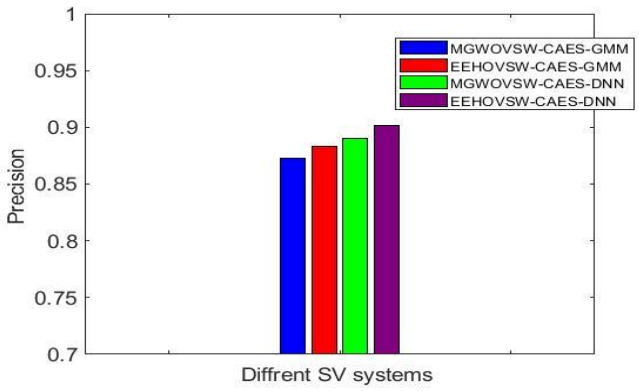

Figure 3. Precision of SV System using Different Algorithms

Figure 3 displays the precision of MGWOVSW-CAES-GMM, EEHOVSW-CAES-GMM, MGWOVSW-CAES-DNN and EEHOVSW-CAES-DNN. It notices that the precision of EEHOVSW-CAES-DNN is 1.27% greater compared to the MGWOVSW-CAES-DNN, 2.11% higher than EEHOVSW-CAES-GMM and 3.37% higher than MGWOVSW-CAES-GMM systems.

4.2 Recall

It is the ratio that measures the total relevant features that are correctly classified.

$$Recall = \frac{TP}{TP + False\ Negative\ (FN)}$$

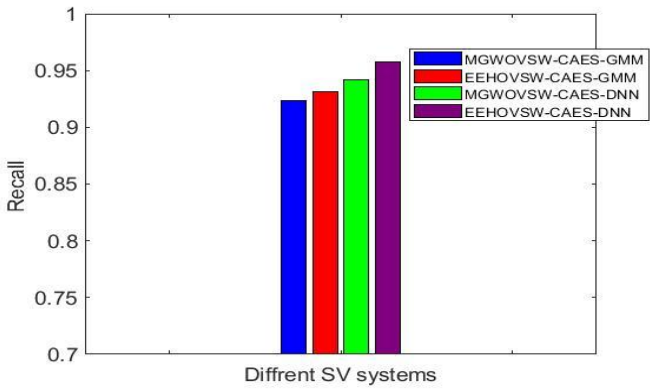

Figure 4. Recall of SV System using Different Algorithms

Figure 4 depicts the recall of MGWOVSW-CAES-GMM, EEHOVSW-CAES-GMM, MGWOVSW-CAES-DNN and EEHOVSW-CAES-DNN. Through this analysis, it is observed that the recall of EEHOVSW-CAES-DNN is 1.63% improved than MGWOVSW-CAES-DNN, 2.81% improved than EEHOVSW-CAES-GMM and 3.66% improved than MGWOVSW-CAES-GMM systems.

4.3 F-measure

It is the harmonic average of precision and recall.

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

Figure 5 portrays the f-measure of MGWOVSW-CAES-GMM, EEHOVSW-CAES-GMM, MGWOVSW-CAES-DNN and EEHOVSW-CAES-DNN. It addresses that the f-measure of EEHOVSW-CAES-DNN is 1.45% increased compared to the MGWOVSW-CAES-DNN, 2.47% higher than EEHOVSW-CAES-GMM and 3.52% higher than MGWOVSW-CAES-GMM systems.

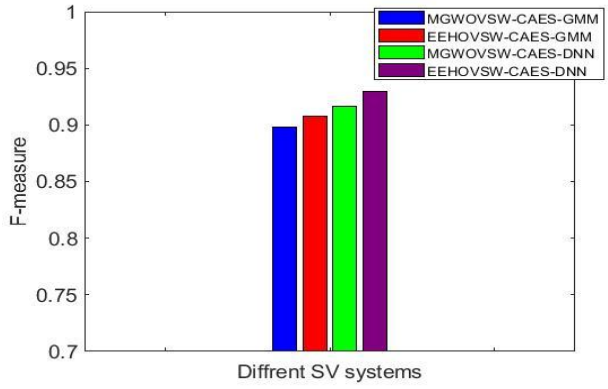

Figure 5. F-measure of SV System using Different Algorithms

4.4 Accuracy

It is the fraction of accurate classification of speech signal features over the total amount of trails examined.

$$Accuracy = \frac{TP + True\ Negative\ (TN)}{TP + TN + False\ Positive\ (FP) + FN}$$

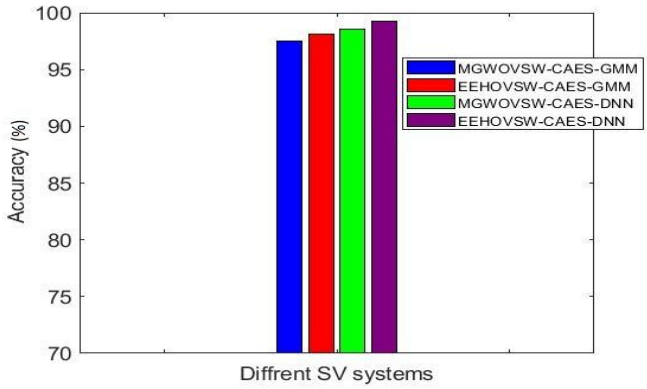

Figure 6. Accuracy of SV System using Different Algorithms

Figure 6 displays the accuracy of MGWOVSW-CAES-GMM, EEHOVSW-CAES-GMM, MGWOVSW-CAES-DNN and EEHOVSW-CAES-DNN. It indicates that the accuracy of EEHOVSW-CAES-DNN is 0.71% greater compared to the MGWOVSW-CAES-DNN, 1.22% higher than EEHOVSW-CAES-GMM and 1.85% higher than MGWOVSW-CAES-GMM systems.

V. CONCLUSION

In this paper, an EEHOVSW-CAES-GMM system is initially proposed that uses the EEHO algorithm for optimizing the speech signal features and parameters of GMM classifier. This EEHO algorithm can update the group and partition functions for achieving the fast convergence and exploitation abilities during training process. But, the GMM classifier has high computation complexity while increasing the number of instances. As a result, MGWOVSW-CAES-DNN and EEHOVSW-CAES-DNN systems are proposed that can minimize the classification errors and the computational complexity if more number of features is extracted and encrypted. In these systems, DNN classifier is applied instead of GMM classifier. Also, the DNNs parameters are optimized by using MGWO and EEHO algorithms. After training the DNN classifier using the decrypted speech signal features, the testing samples are given to verify the customer's individualities. At last, the findings proved the EEHOVSW-CAES-DNN realizes a better authentication efficacy than the MGWOVSW-CAES-DNN and MGWOVSW-CAES-GMM systems.

REFERENCES

[1] May, T., Van De Par, S., & Kohlrausch, A. (2011). Noise-robust speaker recognition combining missing data techniques and universal background modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 108-121.

[2] Bansod, N. S., Kawathekar, S., & Dabhade, S. B. (2012). Review of different techniques for speaker recognition system. *Advances in Computational Research*, *4*(1), 57-60.

[3] Kim, C., & Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *24*(7), 1315-1329.

[4] Indumathi, A., & Chandra, E. (2016). An Efficient speaker recognition system by employing BWT and ELM. *BVICA M's International Journal of Information Technology*, *8*(2), 983-989.

[5] Sahidullah, M., &Kinnunen, T. (2016). Local spectral variability features for speaker verification. *Digital Signal Processing*, *50*, 1-11.

[6] Sreedharan, S., & Eswaran, C. (2018). Optimized variable size windowing based speaker verification. In *ACM Proceedings of the 2018 International Conference on Electronics and Electrical Engineering Technology*, pp. 202-206.

[7] Sujiya Sreedharan & Chandra Eswaran, (2020). Speech feature extraction using RC4 key based AES for speaker verification. *Wulfenia Journal*, *27*, 19-36.

[8] Sujiya Sreedharan, Chandra Eswaran, (2021). A lightweight encryption scheme using chebyshev polynomial maps. *Optik*, *240*, 2-16.

[9] Ramadan, N., Ahmed, H. E. H., Elkhamy, S. E., & El-Samie, F. E. A. (2016). Chaos-based image encryption using an improved quadratic chaotic map. *American Journal of Signal Processing*, *6*(1), 1-13.

[10] Novoselov, S., Pekhovsky, T., Kudashev, O., Mendelev, V. S., & Prudnikov, A. (2015). Non-linear PLDA for i-vector speaker verification. In *Sixteenth Annual Conference of the International Speech Communication Association*, pp. 214-218.

[11] Soleymanpour, M., & Marvi, H. (2017). Text-independent speaker identification based on selection of the most similar feature vectors. *International Journal of Speech Technology*, *20*(1), 99-108.

[12] Thullier, F., Bouchard, B., & Menelas, B. A. (2017). A text-independent speaker authentication system for mobile devices. *Cryptography*, *1*(3), 1-22.

[13] Dey, S., Motlicek, P., Madikeri, S., &Ferras, M. (2017). Template-matching for text-dependent speaker verification. *Speech communication*, *88*, 96-105.

[14] Chowdhury, F. R. R., Wang, Q., Moreno, I. L., & Wan, L. (2018). Attention-based models for text-dependent speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5359-5363.

[15] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnnembeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5329-5333.

[16] Dey, S., Madikeri, S. R., & Motlicek, P. (2018). End-to-end text-dependent speaker verification using novel distance measures. In *Interspeech*, pp. 3598-3602.

[17] Meng, Z., Zhao, Y., Li, J., & Gong, Y. (2019). Adversarial speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6216-6220.

[18] Gupta, M., Bharti, S. S., & Agarwal, S. (2019). Gender-based speaker recognition from speech signals using GMM model. *Modern Physics Letters B*, *33*(35), 1-23.

[19] Abualadas, F. E., Zeki, A. M., Al-Ani, M. S., & Messikh, A. E. (2019). Speaker identification based on hybrid feature extraction techniques. *International Journal of Advanced Computer Science and Applications*, *10*(3), 322-327.

[20] Shon, S., Tang, H., & Glass, J. (2019). Voiceid loss: Speech enhancement for speaker verification. In *Interspeech*, pp. 2888-2892.

[21] Liu, Y., He, L., Liu, J., & Johnson, M. T. (2019). Introducing phonetic information to speaker embedding for speaker verification. *EURASIP Journal on Audio, Speech, and Music Processing*, *2019*(1), 1-17.

[22] Kakade, M. N., & Salunke, D. B. (2020). An automatic real time speech-speaker recognition system: a real time approach. In *Proceedings of the 2nd International Conference on Communications and Cyber Physical Engineering*, Springer, Singapore, pp. 151-158.

[23] Arora, S. V., & Vig, R. (2020). An efficient text-independent speaker verification for short utterance data from Mobile devices. *Multimedia Tools and Applications*, *79*(3), 3049-3074.

[24] Sun, R. H., & Chol, R. J. (2020). Subspace Gaussian mixture based language modeling for large vocabulary continuous speech recognition. *Speech Communication*, *117*, 21-27.

[25] Ramteke, P. B., Supanekar, S., & Koolagudi, S. G. (2020). Classification of aspirated and unaspirated sounds in speech using excitation and signal level information. *Computer Speech & Language*, 1-18.

Funding: The research work is supported by Department of Science and Technology-PURSE (Phase – II). This work has been submitted for Indian Intellectual property with Patent Application Number 201841032393

Conflicts of interest/Competing interests
The authors have declared no conflict of interest

Corresponding author

Sujiya Sreedharan

Availability of data and material: Not applicable

Code availability: Not applicable

Figures

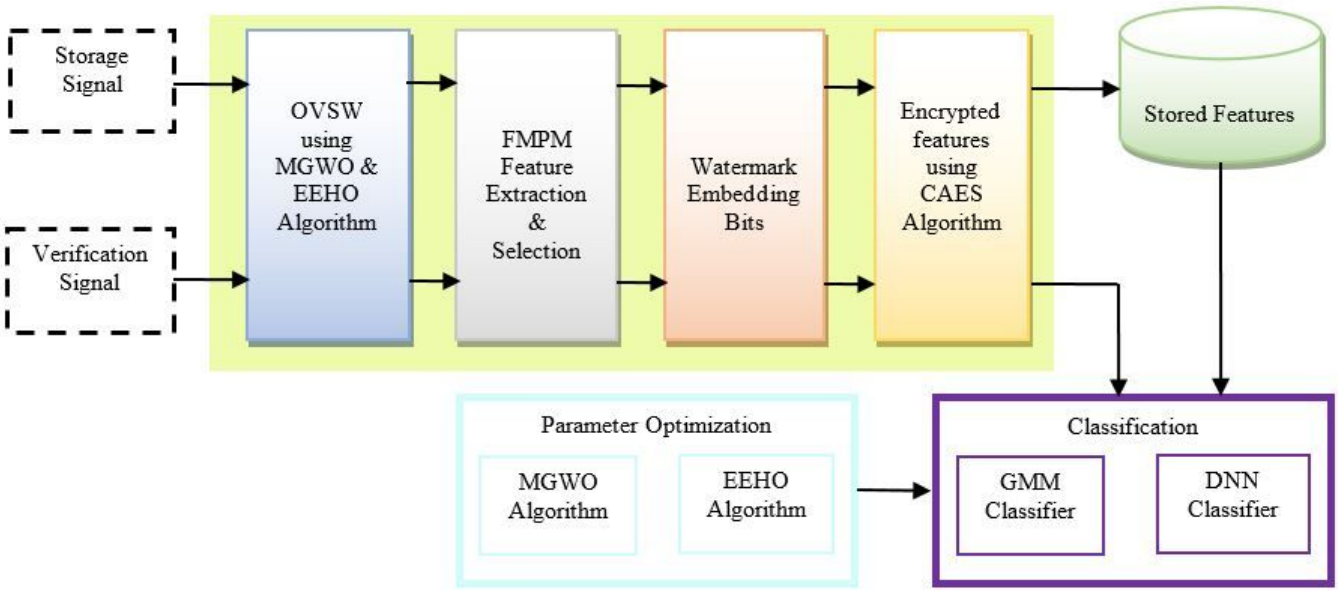

Figure 1

Schematic Representation of Proposed OVSW-SV Systems using Two Different Optimization Algorithms

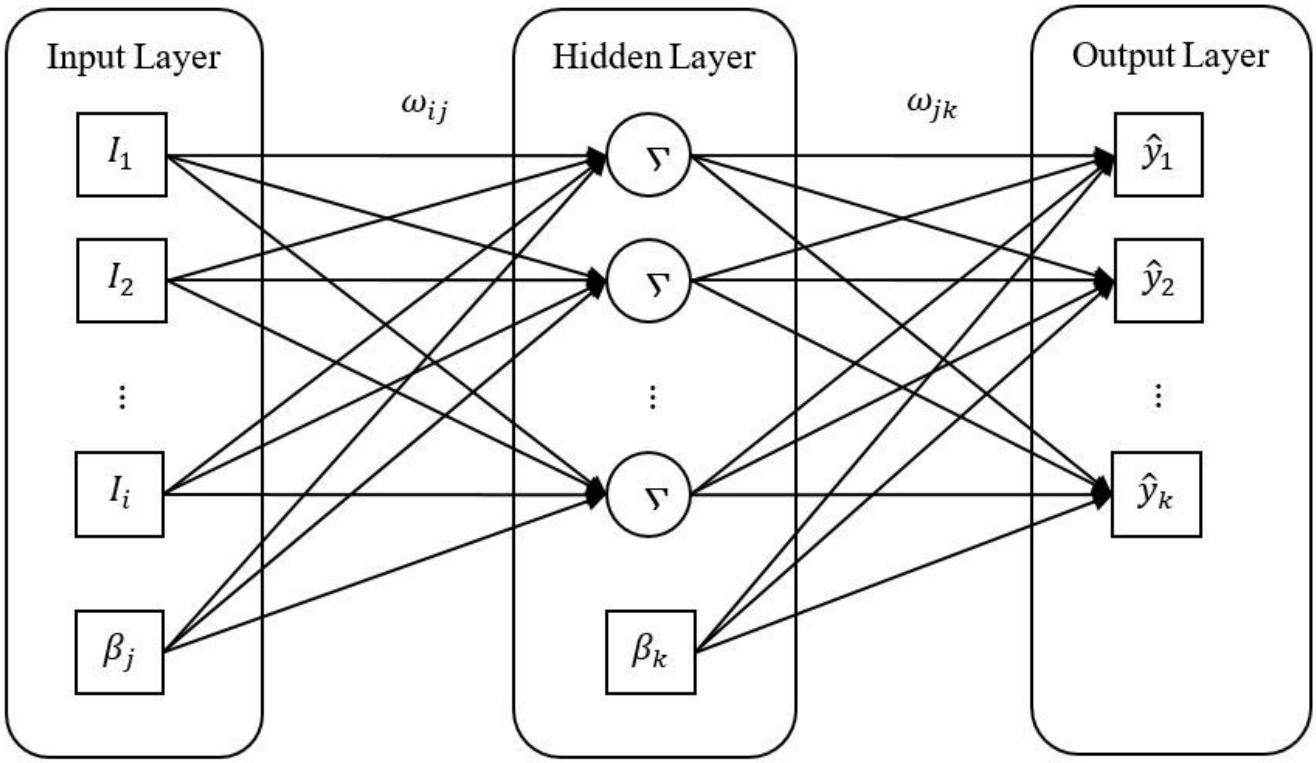

Figure 2

Schematic Structure of DNN Classifier

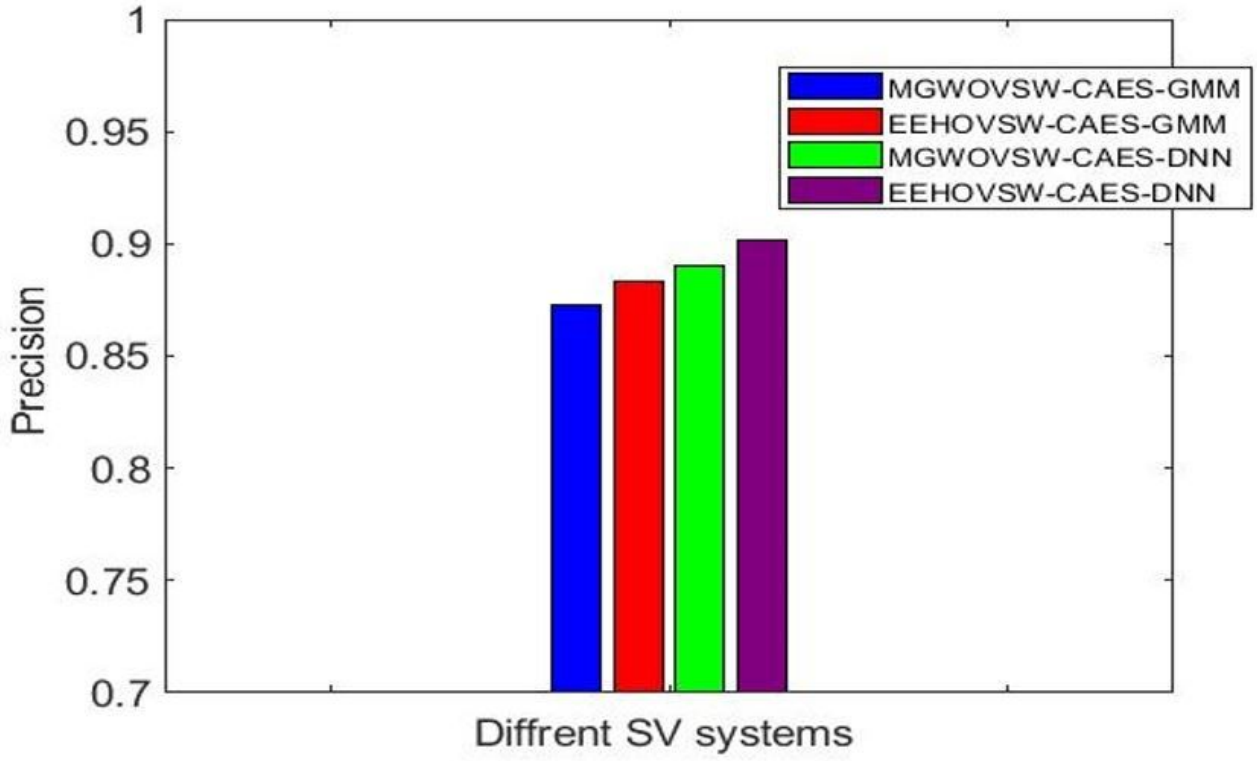

Figure 3

Precision of SV System using Different Algorithms

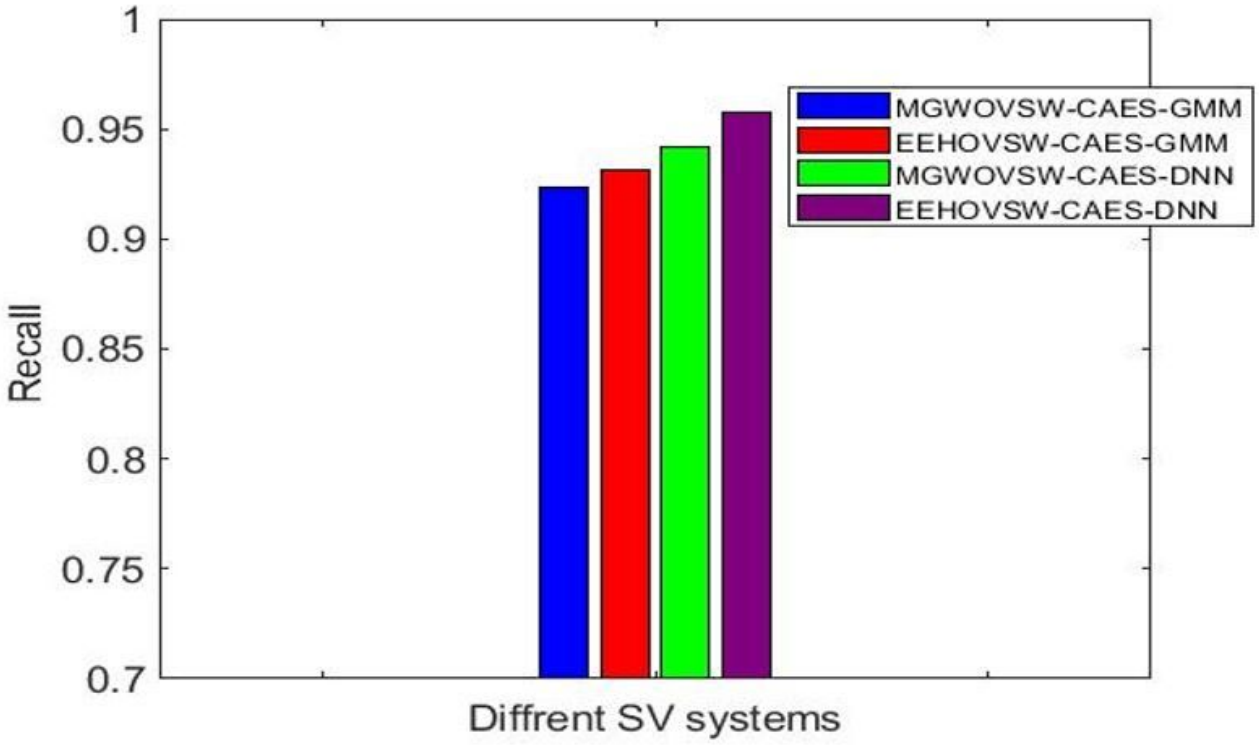

Figure 4

Recall of SV System using Different Algorithms

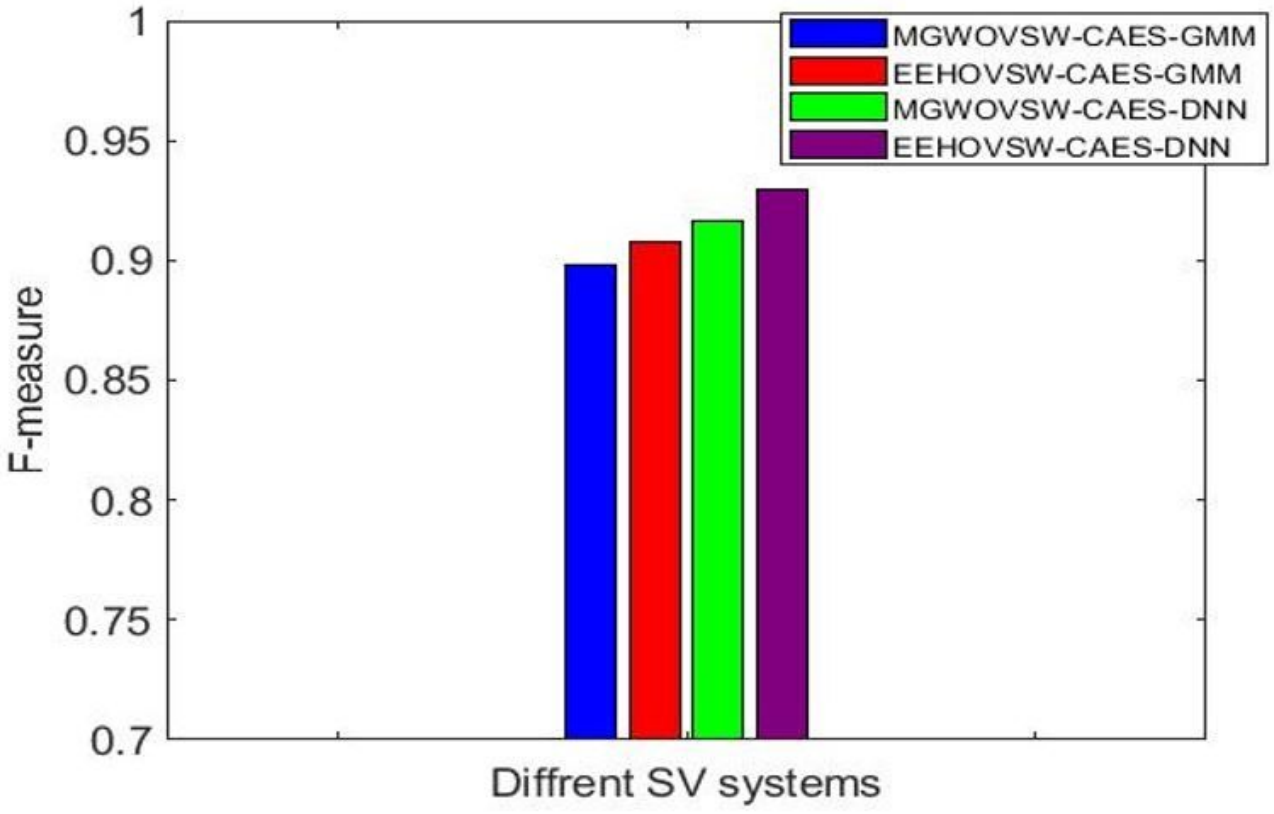

Figure 5

F-measure of SV System using Different Algorithms

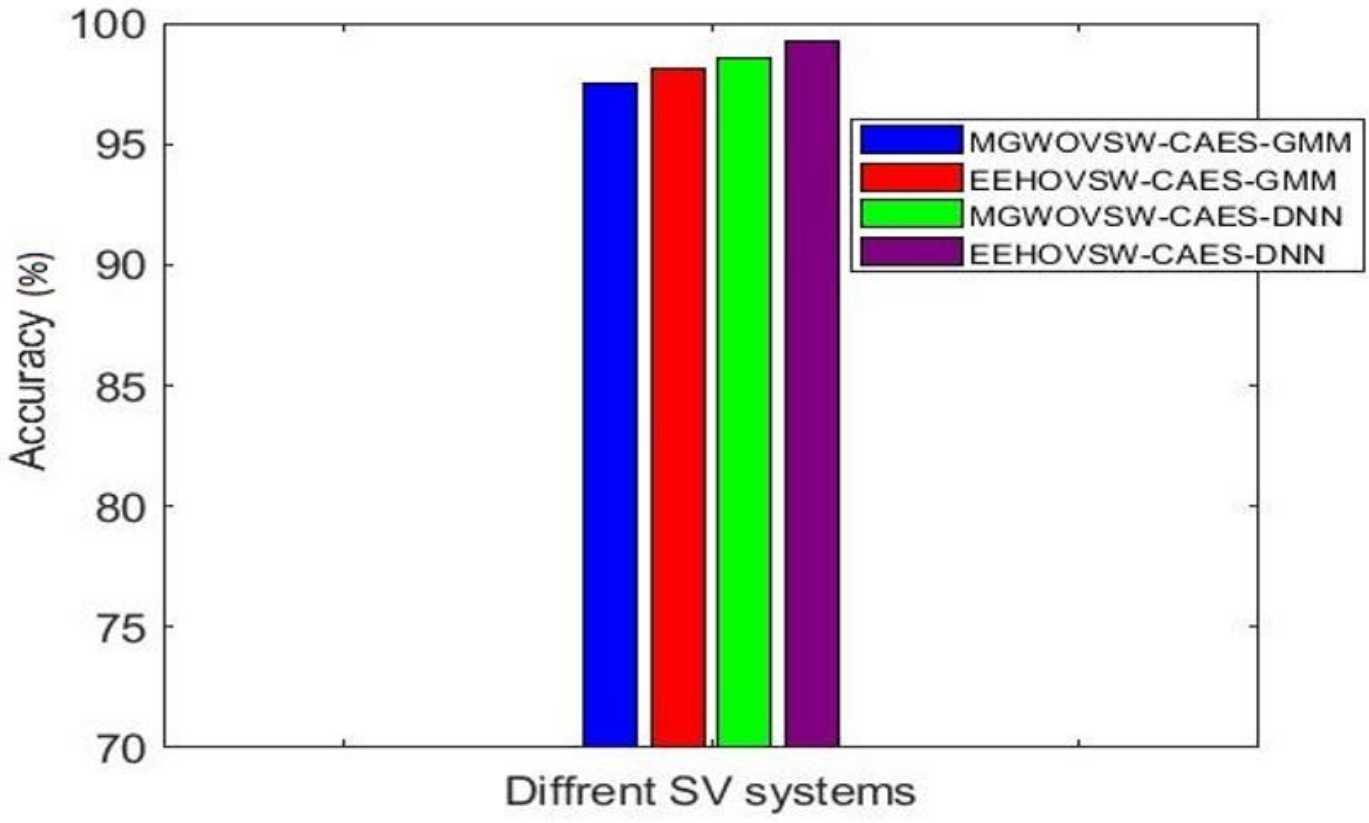

Figure 6

Accuracy of SV System using Different Algorithms