LAUTECH Journal of Engineering and Technology 18 (4) 2024: 129-139
10.36108/laujet/4202.81.0431

Development of Speech Emotion Recognition System Using Optimized Convolutional Neural Network

[1*]Adebiyi B. F., [2] Oke A. O., [3] Falohun A. S. and [4]Awodoye O. O.

[1,2,3,4]Department of Computer Engineering, Ladoke Akintola University of Technology, Ogbomoso

[1]adebiyibo@gmail.com , [2]aooke@lautech.edu.ng , [3]asfalohun@lautech.edu.ng, [4]ooawodoye50@lautech.edu.ng

Article Info

Article history:

Received: Dec. 31, 2024
Revised: Jan. 16, 2025
Accepted: Jan. 19, 2025

Keywords:

Mantis Search Algorithm,
Speech Emotion Recognition,
Convolutional Neural Network,
Natural Language Processing

Corresponding Author:
adebiyibo@gmail.com

ABSTRACT

Speech Emotion Recognition (SER) enables systems to interpret emotions in human speech, facilitating natural human-machine interactions. Despite significant advancements in SER using optimization algorithms, existing methods often faced challenges with convergence speed, computational efficiency, and balancing exploration and exploitation in high-dimensional parameter spaces. Due to the complex nature of emotion detection, several deep learning techniques have been utilized, yet limited studies have focused on optimizing key hyperparameters of the Convolutional Neural Network (CNN) for a more efficient system. Hence, this research optimized CNN with Mantis Search Algorithm (MSA) due to its simplicity, ability to maintain diversity, escape local optima, and balance exploration and exploitation. Audio data for anger, fear, happiness, and neutrality were acquired from the Toronto Emotional Speech Set (TESS) on Kaggle.com. The data were converted to text using speech-to-text code and preprocessed with Natural Language Processing (NLP) techniques - tokenization, stop-word removal, lemmatization, punctuation removal, and lowercase conversion. MSA optimized CNN by selecting the optimal filter size and learning rate. The resulting MSA-CNN was implemented in MATLAB R2023a. The performance of the system was evaluated and compared with the CNN classifier using False Positive Rate (FPR), Specificity (Spec), Sensitivity (Sen), Precision (Prec), Accuracy (Acc), and Recognition Time (RT). On average, MSA-CNN achieved lower FPR (0.70% compared to 1.61%), higher Specificity (99.30% compared to 98.40%), greater Sensitivity (97.28% compared to 94.55%), improved Precision (97.89% compared to 95.15%), better Accuracy (98.80% compared to 97.43%), and reduced Recognition Time (68.32s compared to 96.87s).

INTRODUCTION

Humans interact with machines like smartphones and advanced gadgets for their ease, speed, and accuracy. This effective communication is called Human-Computer Interaction (HCI) (Li and Jibrin, 2016; Adetunji *et al.*, 2018). HCI offers a large scope of accessibility even to people with impairment and disorder as it provides a way to communicate with speech in cases of hand disabilities or sight restraints (Manjunath *et al*, 2022). HCI also enhances human development and online learning improvement. A highly significant part of the world's population communicates through spoken words (speech) as it provides an incredibly fast and easy way of communication. Speech is a natural and widely used method of human communication, conveying both linguistic and paralinguistic information.

Linguistic information pertains to the content and language of the speech, while paralinguistic information reflects characteristics such as gender, emotions, age, and other unique human traits (Ala *et al*, 2023). Emotions form an integral part of human interactions and have naturally become an important

aspect of the development of HCI-based applications. Emotions can be captured and analyzed through various technologies, including facial expressions, physiological signals, and speech. Speech can convey a variety of emotions, including anger, pride, boredom, disgust, surprise, fear, joy, happiness, neutrality, and sadness (Ruhul *et al*, 2019).

These emotions can be detected in speech through Speech Emotion Recognition (SER), which involves using Artificial Intelligence (AI) techniques to predict human emotions from audio signals (Ala *et al*, 2023). SER is a key feature in HCI and may be considered a branch of Automatic Speech Recognition (ASR) exploiting the same kind of signal, feature extraction processes, and potential application of diverse machine learning techniques. These include Deep Learning (DL) architectures like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), which are also applied in the field of Natural Language Processing (NLP) (Javier and Manuel, 2023; Oguntoye *et al*., 2023; Atanda *et al*., 2023).

NLP is most directly associated with processing human (natural) language. It integrates computational linguistics with rule-based language modeling, alongside statistical, machine learning, and deep learning techniques. These combined technologies enable computers to process human language, whether in text or voice and comprehend its full meaning, including the speaker's or writer's intent and sentiment (Trilla, 2009). NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly even in real-time using several tools and methods.

Mantis Search Algorithm (MSA), a bio-inspired algorithm was adjudged efficient and effective for global optimization and engineering design problems (Mohamed *et al*, 2023). MSA is an optimization method inspired by nature that reflects the hunting techniques and mating behaviors of praying mantises. Mantises are distinguished by their elongated bodies, triangular heads equipped with compound eyes, and flexible necks. Mantises employ various hunting strategies, including camouflage, ambush tactics, and striking with modified forelegs. These insects feed on a range of prey, including wasps, spiders, ants, and even small vertebrates.

MSA comprises three key stages: the exploration phase, where it imitates a mantis searching for prey, the exploitation phase, which emulates a mantis attacking its target; and sexual cannibalism, a phenomenon in which female mantises attract male mantises to their habitat for mating, only to subsequently devour the male during the process (Mohamed *et al*, 2023). MSA is easy to implement, preserves population diversity during the optimization process, and has a high ability to escape from the local optima and balance between exploration and exploitation operators. SER technologies have a wide range of applications in areas such as psychology, medicine, education, and entertainment (Inna and Daria, 2020).

LITERATURE REVIEW

Wani *et al*, (2020) introduced a CNN—Stride-based Convolutional Neural Network (SCNN) model that employed fewer convolutional layers and removed pooling layers to enhance computational stability. This approach typically enhanced accuracy while reducing the computational time of the SER system. Instead of pooling layers, deep strides were used for the necessary dimension reduction. The SCNN model was trained on spectrograms generated from the speech signals of two different databases, Berlin (Emo-DB) and IITKGP-SEHSC. Four emotions, anger, happiness, neutral, and sad, were considered

for the evaluation process, and a validation accuracy of 90.67% and 91.33% was achieved for Emo-DB and IITKGP-SEHSC, respectively.

Wen *et al*, (2021) proposed a new machine-learning model by fusing CNN and Capsule Network (CapsNet) called CapsCNN for speech emotion recognition. The datasets used were CASIA and EmoDB. The extracted features for both datasets included MFCC and spectrogram. The result of the experiment showed that CapsCNN has higher accuracy above 80% for both datasets when compared to CNN.

Fazliddin *et al*, (2022) proposed a novel SER model through attention-oriented parallel CNN encoders that parallelly acquired important features used for emotion classification. Specifically, MFCC, paralinguistic, and speech spectrogram features were extracted and encoded by developing distinct CNN architectures tailored for each feature. The encoded features were then input into attention mechanisms for enhanced representation before being classified. The EMO-DB and IEMOCAP open datasets were utilized, and the findings demonstrated that the proposed model surpassed the baseline models in efficiency. In the context of the EMO-DB dataset, the proposed model recorded a Weighted Accuracy (WA) of 71.8% and an Unweighted Accuracy (UA) of 70.9%.

Yunhao and Xiaoqing (2023), presented speech emotion analysis using CNN and Gamma classifier-based Error Correcting Output Codes (ECOC) to improve the performance of emotion analysis in speech. CNN was used to reduce the dimensionality of these features and extract the characteristics of each signal. A combination of the Gamma Classifier (GC) and ECOC was used to classify the features and identify emotions (anger, hatred, joy, fear, and sadness) in speech. The performance of the method was evaluated using two datasets, Berlin and ShEMO. The results showed that the proposed method achieved average accuracies of 93.33% and 85.73% in recognizing speech emotions on the Berlin and ShEMO datasets, respectively.

Islam *et al*, (2024) proposed an enhanced speech-emotion recognition system using Deep Convolutional Neural Networks (DCNNs) for the enhancement of accuracy and precision of emotion prediction to improve emotion prediction's precision and accuracy. The Toronto Emotional Speech Set (TESS) database, which included seven different emotions—disgust, fear, rage, happiness, sorrow, and neutrality—was used by the authors and 88% accuracy was achieved.

From the aforementioned reviewed works, it was discovered that MSA as effective as it is has not been used for optimization in the area od SER. Hence, this research developed an SER system using CNN optimized with Mantis Search Algorithm (MSA) for CNN's hyperparameters optimization using False Positive Rate, Specificity, Sensitivity, Precision, accuracy, and Recognition Time as evaluation metrics.

METHODOLOGY

The developed system is based on the stages: audio acquisition, speech-to-text conversion, text preprocessing, feature extraction and classification, and result evaluation as shown in Figure 1.

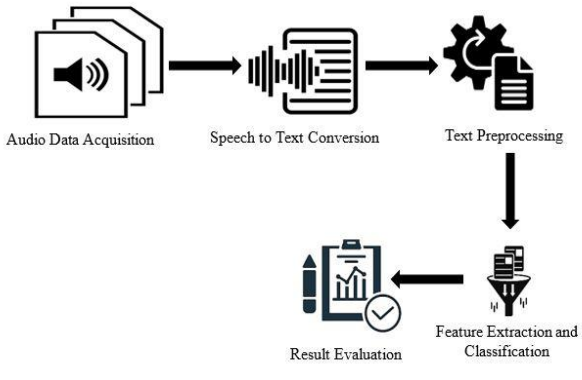

Figure 1: The Stages of the Developed System

Audio Data Acquisition

The dataset used was the Toronto Emotional Speech Set (TESS) acquired from Kaggle. This dataset includes audio records in .wav file format for seven emotions, of which four emotions: anger, fear, happiness and neutral were chosen. Anger, fear, and happiness were chosen as they represent basic emotions, while neutral was included as a baseline emotion. The audio records totaled 6368 with 1584 records for anger, 1588 for fear, 1596 for happiness and 1600 for neutral emotions.

Speech-to-Text Conversion

The audio records were converted into text using the speech-to-text code and the output was saved in an Excel Comma Separated Values (CSV) file format. The CSV file had columns for each of the corresponding text, score and emotion as shown in Table 1. Neutral had a score of 1, happiness had a score of 2, fear had a score of 3 and anger had a score of 4.

Table 1: Emotion and corresponding score

Emotion	Score
Neutral	1
Happiness	2
Fear	3
Anger	4

Text Preprocessing

The texts obtained from the audio records were preprocessed using Natural Language Processing (NLP) techniques to remove parts and constituents that were not needed for analysis. Letters of words present in each text were converted to lowercase for consistency and text improvement. Then, punctuations were removed from the texts. Tokenization was used to break down and segment each text into smaller or individual pieces called tokens for easy analysis of individual words.

Stop words were removed from the texts. Stop words are frequently used or common words that do not carry significant meaning or value in a text. Stop words include articles, prepositions, pronouns, and auxiliary verbs. These stop words are not required for emotional analysis of the text. Hence, such words were discarded so that the quality of the text was improved. Removal of stop words was used to reduce noise and increase the efficiency of the texts.

Lemmatization was used to obtain the root or base word of every word in each text to know the intended meaning of each word. Root words are meaningful base forms of the text, called the lemma. Lemmatization normalized words in the text thereby reducing the dimensionality of each text for easy capturing of information carried by each text. The final output of the preprocessing stage was then converted to a sequence of numerical indices. An example of the preprocessing stage is shown in Figure 2.

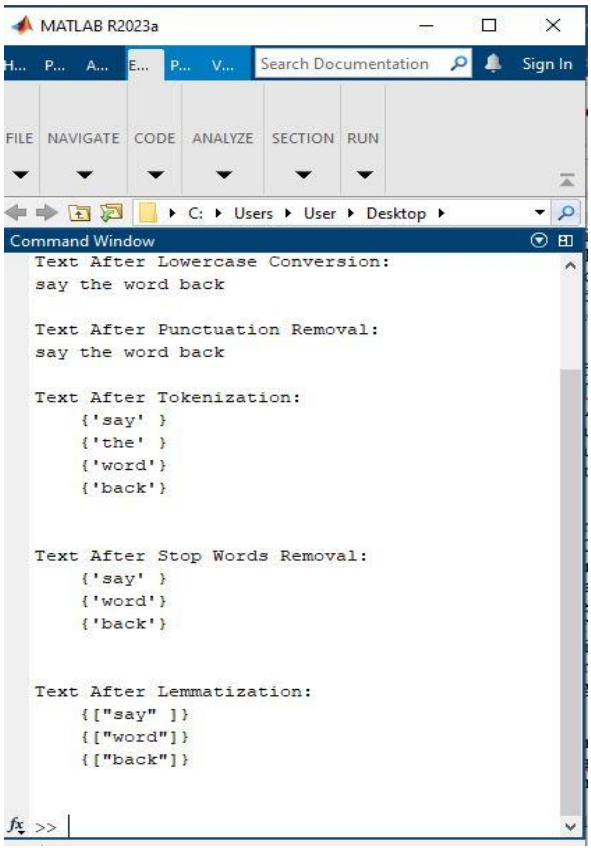

Figure 2: Text Preprocessing

Feature Extraction and Classification

Convolution Neural Network (CNN) was used as a deep learning technique for the research. The standard CNN algorithm is as shown in Algorithm 1. The sequence of numerical indices served as input to train the CNN model. At the feature extraction phase of CNN, the convolution layer was optimized. This optimization was achieved by using the Mantis Search Algorithm (MSA) as shown in Algorithm 2 to select the optimal value for filter size and learning rate. Filter size is the dimension of filters or kernels used in convolutional layers to detect specific features and patterns.

Algorithm 1: Standard CNN Algorithm

1	Initialize the network parameters
2	Input the data into the network
3	Perform a convolution operation with multiple filters to extract feature maps
4	Apply a non-linear activation function to the feature maps.
5	Perform pooling operations to reduce the spatial dimensions of the feature maps.
6	Repeat steps 3-5 for several convolutional and pooling layers to build deeper feature representations.
7	Flatten the output of the final pooling layer to create a 1D feature vector.
8	Pass the flattened feature vector through fully connected (dense) layers.
9	Apply a non-linear activation function to the outputs of the fully connected layers.
10	Apply an output layer for the final classification or regression task.
11	Compute the loss using an appropriate loss function.
12	Perform backpropagation to compute gradients of the loss concerning the network parameters.
13	Update the network parameters using an optimization algorithm (e.g., gradient descent, Adam).
14	Repeat steps 2-13 for multiple epochs until the network is trained.
15	Evaluate the trained network on validation/test data to assess

Algorithm 2: MSA-CNN Algorithm

1	Initialize Population: Randomly initialize the population with different sets of hyperparameters (filter size, learning rate).
2	For each set of hyperparameters in the population, train the CNN and evaluate its performance (accuracy on a validation set).
3	Select the top-performing sets of hyperparameters based on their fitness scores to form a mating pool.
4	Exploration Stage Adjust the hyperparameters to explore the search space more broadly. This helps in identifying new potential solutions.
5	Exploitation Stage Focus on refining the solutions around the best-performing individuals to

	intensify the search in promising areas.
6	Sexual Cannibalism Weaker individuals are removed or significantly altered to introduce diversity and avoid local optima.
7	Generate new individuals by mating the selected ones and applying crossover and mutation to their hyperparameters.
8	Update the Population by replacing the old population with the new offspring and adjusted individuals.
9	Check if the termination criteria are met.
10	Return the best set of hyperparameters found as the optimal result.
11	Initialize the CNN parameters (e.g., weights and biases) using the optimal hyperparameters obtained from MSA.
12	Input the data into the CNN.
13	Perform a convolution operation with the optimized filter size to extract feature maps.
14	Apply a non-linear activation function (ReLU) to the feature maps.
15	Perform pooling operations (e.g., max pooling) to reduce the spatial dimensions of the feature maps.
16	Repeat steps 13-15 for several convolutional and pooling layers to build deeper feature representations.
17	Flatten the output of the final pooling layer to create a 1D feature vector.
18	Pass the flattened feature vector through fully connected (dense) layers.
19	Apply a non-linear activation function to the outputs of the fully connected layers.
20	Output Layer: Apply the output layer (softmax) for the final classification or regression task.
21	Compute the loss using the loss function.
22	Perform backpropagation to compute gradients of the loss concerning the network parameters.
23	Update the network parameters with the optimized learning rate.
24	Repeat steps 12-23 for multiple epochs until the network is trained.
25	Evaluate the trained network on validation/test data to assess its performance.

The learning rate is a CNN hyperparameter that controls how quickly a system learns from the training data (Olayiwola *et al*., 2023). MSA has three phases: the exploration phase, the exploitation phase and the sexual cannibalism phase. From Algorithm 2, at the exploration stage, the algorithm searched through random combinations of the hyperparameters to identify the optimal combination that maximizes the fitness function. Then, at the exploitation stage, the best set of combinations of the hyperparameters was selected. The weaker combinations of the hyperparameters were replaced

with the stronger ones to perfect the best combinations. At the sexual cannibalism stage, the best combination was chosen to improve the output of computation.

Evaluation Metrics

The performance of this research was evaluated based on false positive rate, specificity, sensitivity, precision, accuracy and recognition time.

False Positive Rate (FPR)

This is calculated as the ratio of false positives to the total number of actual negatives. It is important to minimize the false positive rate to ensure accurate identification of the intended emotions.

$$\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$$

Precision

This indicates the proportion of correctly identified target emotions out of all the instances labeled as the target emotion by the system.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

Accuracy

This measures the overall correctness of the system's predictions across all classes of emotions. accuracy shows the proportion of correctly classified instances (both positive and negative) out of all the instances in the dataset.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+ FP+FN}}$$

Recognition Time

This refers to the duration it takes for the system to process and analyze an input speech signal and provide an output emotion classification. It measures the speed or efficiency of the speech-emotion recognition system. Recognition time is often considered alongside other metrics to evaluate the overall performance of the speech-emotion recognition system.

Where:

i. FP denotes the number of false positives, i.e., instances where the system incorrectly identifies a non-target emotion as the target emotion.
ii. TN denotes the number of true negatives, i.e., instances where the system correctly identifies non-target emotions as non-target emotions.
iii. TP represents the number of true positives, i.e., instances where the system correctly identifies the target emotion.
iv. FN represents the number of false negatives, i.e., instances where the system fails to identify the target emotion when it is present.

RESULTS AND DISCUSSION

The total data used in this research was 6368 with 4458 (70%) used for training and 1910 (30%) used for testing using random sampling cross-validation method.

Results for CNN

The result of the CNN technique based on 1910 audio data used for testing at an optimum threshold of 0.8 are shown in Table 2. 475, 477, 478 and 480 audio data were for anger, fear, happiness and neutral emotion, respectively. For emotion anger, 449 audio data were correctly classified with 26 audio data misclassified. As a result of this, the classifier had FPR of 1.60%, Spec of 98.40%, Sen of 94.53%, Prec of 95.12%, Acc of 97.43% and RT of 92.44s.

For emotion fear, 451 audio data were correctly classified as fear with 26 audio data misclassified. As a result of this, the classifier had FPR of 1.61%, Spec of 98.39%, Sen of 94.55%, Prec of 95.15%, Acc of 97.43% and RT of 94.94s. For happiness, 452 audio data were correctly classified with 26 audio data misclassified. As a result of this, the classifier

had FPR of 1.61%, Spec of 98.39%, Sen of 94.56%, Prec of 95.16%, Acc of 97.43% and RT of 97.89s. For neutral emotion, 454 audio data were correctly classified as fear with 26 audio data misclassified. As a result of this, the classifier had FPR of 1.61%, Spec of 98.39%, Sen of 94.58%, Prec of 95.18%, Acc of 97.43% and RT of 100.22s.

Table 2: CNN Result at Optimum Threshold

Emotion	Anger	Fear	Happiness	Neutral
TP	449	451	452	454
FN	26	26	26	26
FP	23	23	23	23
TN	1412	1410	1409	1407
FPR(%)	1.60	1.61	1.61	1.61
Spec(%)	98.40	98.39	98.39	98.39
Sen(%)	94.53	94.55	94.56	94.58
Prec(%)	95.13	95.15	95.16	95.18
Acc(%)	97.43	97.43	97.43	97.43
Time(sec)	92.44	94.94	97.89	100.22

Result for MSA-CNN

The result of the MSA-CNN technique based on 1910 audio data used for testing at an optimum threshold of 0.8 are shown in Table 3. 475, 477, 478 and 480 audio data were for anger, fear, happiness and neutral emotion, respectively. For anger, 462 audio data were correctly classified with 13 audio data misclassified. As a result of this, the classifier had FPR of 0.70%, Spec of 99.30%, Sen of 97.26%, Prec of 97.88%, Acc of 98.80% and RT of 62.24s. For fear, 464 audio data were correctly classified with 13 audio data misclassified. As a result of this, the classifier had FPR of 0.70%, Spec of 99.30%, Sen of 97.27%, Prec of 97.89%, Acc of 98.80% and RT of 67.42s.

For happiness, 465 audio data were correctly classified with 13 audio data misclassified. As a result of this, FPR was 0.70%, Spec was 99.30%, Sen was 97.28%, Prec was 97.89%, Acc was 98.80% and RT was 72.64. For neutral emotion, 467 audio data were correctly classified with 13 audio data misclassified. As a result of this, the classifier had FPR of 0.70%, Spec of 99.30%, Sen of 97.29%, Prec of 97.90%, Acc of 98.80% and RT of 72.98s.

Table 3: MSA-CNN Result at Optimum Threshold

Emotion	Anger	Fear	Happiness	Neutral
TP	462	464	465	467
FN	13	13	13	13
FP	10	10	10	10
TN	1425	1423	1422	1420
FPR(%)	0.70	0.70	0.70	0.70
Spec(%)	99.30	99.30	99.30	99.30
Sen(%)	97.26	97.27	97.28	97.29
Prec(%)	97.88	97.89	97.89	97.90
Acc(%)	98.80	98.80	98.80	98.80
Time(sec)	62.24	67.42	72.64	72.98

Comparison of CNN and MSA-CNN Results

The comparative analysis of CNN and MSA-CNN for the performance metrics at an optimum threshold of 0.8 are shown in Figures 3, 4, 5 and 6. For all emotion classification, when compared to CNN, the MSA-CNN classifier demonstrated a robust performance profile, exhibiting a combination of lower FPR, higher specificity, higher sensitivity, higher precision, higher accuracy, and reduced recognition time. These characteristics collectively indicate that the MSA-CNN classifier is more

accurate and efficient, with implications for its reliability in various high-stakes applications.

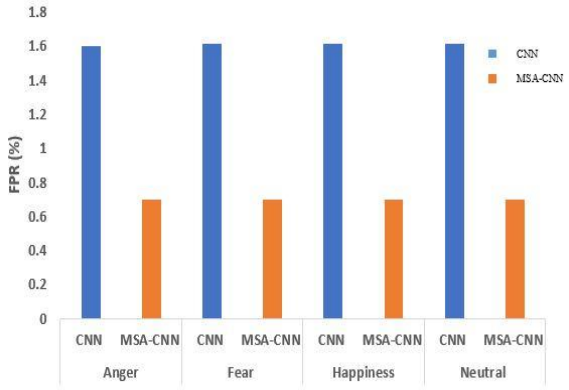

Figure 3: FPR for CNN and MSA-CNN

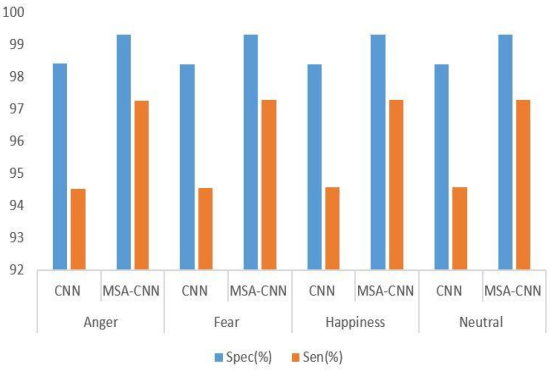

Figure 4: Spec and Sen for CNN and MSA-CNN

The lower FPR shows that the classifier made few false-positive predictions. Higher specificity further enhances reliability by enabling the classifier to effectively identify true negatives. This improves confidence in cases classified as negative, reducing the potential for unwarranted positive predictions and strengthening the model's trustworthiness in applications where distinguishing true negatives is essential.

Higher sensitivity implies that the classifier is highly effective at identifying true positive cases. That is, it captures a large proportion of actual positives, thereby minimizing instances of false negatives. Higher precision means the classifier is more accurate when predicting positive cases, resulting in fewer false positives.

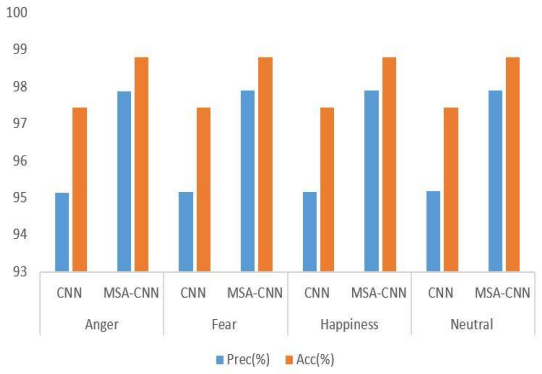

Figure 5: Prec and Acc for CNN and MSA-CNN

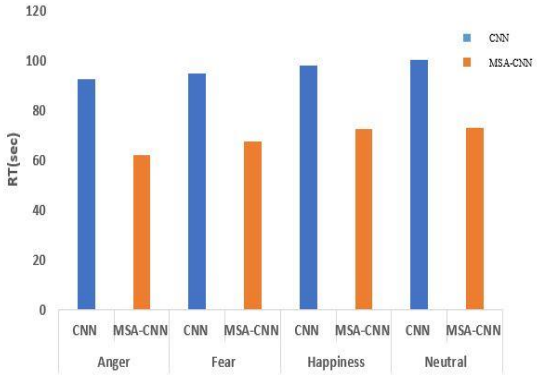

Figure 6: RT for CNN and MSA-CNN

Overall, the classifier's higher accuracy demonstrates a well-balanced performance across both positive and negative classes, indicating that it makes fewer errors on the whole. This balanced accuracy suggests that the model is generalized effectively, performing reliably across a range of cases. Complementing these strengths, the classifier's reduced recognition time highlights its efficiency, enabling it to make predictions quickly.

The classifier's lower FPR, higher specificity, sensitivity, precision, and accuracy, along with reduced recognition time, indicates that it is both more accurate and timely in its predictions compared to the CNN classifier.

CONCLUSION

In this research, speech emotion recognition was implemented using an optimized Convolutional

Neural Network (MSA-CNN). Filter size and learning rate, the hyperparameters of Convolutional Neural Network (CNN) were optimized using the Mantis Search Algorithm (MSA) by using the algorithm to select the optimal values of the chosen hyperparameters. The evaluation performance of CNN and MSA-CNN in speech emotion recognition was determined using False Positive Rate, Specificity, Sensitivity, Precision, accuracy, and Recognition Time as metrics.

The MSA-CNN classifier was discovered to outperform the CNN classifier in terms of all the evaluation metrics used thereby making it more efficient and reliable. The developed system can be applied in customers' service for generating feedback on goods and services. Also, it can be of significant use in medicine for psychological analysis and treatment of patients. The system can also be used to detect hate speech.

REFERENCES

Adetunji A. B., Oguntoye J. P., Fenwa O. D. and Omidiora E. O. (2018): Reducing the Computational Cost of SVM in Face Recognition Application Using Hybrid Cultural Algorithm. IOSR Journal of Computer Engineering (IOSR-JCE). 20 (2): 36-45.

Ala, S. A., Oumaima, S., Rashid, J., Muhammad, A. N., and Omnia, S. N. (2023). Speech emotion recognition through hybrid features and convolutional neural network. *Applied Science.* 13: 1-15.

Atanda, O. G., Ismaila, W., Afolabi, A. O., Awodoye, O. A., Falohun, A. S., & Oguntoye, J. P. (2023). Statistical Analysis of a Deep Learning Based Trimodal Biometric System Using Paired Sampling T-Test. In 2023 International Conference on Science, Engineering and Business for Sustainable Development Goals (SEB-SDG). 1: 1-10.

Fazliddin, M., Alpamis, K., Farkhod, A., Mohamed, S. A., and Young-Im, C. (2022). Modeling speech emotion recognition via attention-oriented parallel CNN encoders. *Electronics*, 11: 1-14.

Inna S. M., and Daria V. T. (2020). Recognition of emotions in verbal messages based on neural networks. *Procedia Computer Science*, 190: 560–563.

Islam, M. M., Kabir, M. A., Sheikh, A., Saiduzzaman, M., Hafid, A., and Abdullah, S. (2024). Enhancing speech emotion recognition using deep convolutional neural networks. Proceedings of the 2024 9th International Conference on Machine Learning Technologies (ICMLT 2024). https://doi.org/10.1145/3674029.3674045: 95-100.

Javier, D., and Manuel, G. (2023). An ongoing review of speech emotion recognition. *Neurocomputing,* 528: 1-11.

Li, P., and Jibrin, Y. (2016). An overview of human performance models in human-computer interactions. *International Conference on Intelligent Control and Computer Application (ICCA)*, 421-425.

Manjunath, H. R., Ananya, H. S., Anusha, T. R., Arundhati, S. B., and Bhagyalakshmi A. N. (2022). Review on human computer interaction. *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, 2(1): 720-726.

Mohamed. A., Reda, M., Mahinda, Z., Mohammed, J., and Mohamed, A. (2023). Mantis Search Algorithm: A novel bio-inspired algorithm for global optimization and engineering design problems. *Computer Methods in Applied Mechanics and Engineering,* 415: 1-43.

Oguntoye, J. P., Awodoye, O. O., Oladunjoye, J. A., Faluyi, B. I., Ajagbe, S. A., & Omidiora, E. O. (2023). Predicting COVID-19 From Chest X-

Ray Images using Optimized Convolution Neural Network. LAUTECH Journal of Engineering and Technology, 17(2): 28-39.

Olayiwola, D. S., Olayiwola, A. A., Oguntoye, J. P., Awodoye, O. O., Ganiyu, R. A., & Omidiora, E. O. (2023). Development of a Fingerprint Verification and Identification System Using a Gravitational Search Algorithm-Optimized Deep Convolutional Neural Network. Adeleke University Journal of Engineering and Technology, 6(2): 296-307.

Ruhul, A. K., Edward, J., Mohammed, I. B., Tariqullah, J., Mohammad, H. Z. and Thamer A. (2019). Speech emotion recognition using deep learning Techniques: a review. *IEEE*, 7: 117327-117345.

Trilla, A., (2009). Natural language processing techniques in text-to-speech synthesis and automatic speech recognition. Working Paper, Enginyeria i Arquitectura La Salle, Universitat Ramon Llull, Barcelona, Spain.

Wani, T. M., Gunawan, T. S., Qadri, S. A., Mansor, H., Arifin, F. and Ahmad, Y. A., (2020). Stride based convolutional neural network for speech emotion recognition. *IEEE 7th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, 1-6.

Wen, X. C., Liu, K. H., Zhang, W. M., and Jiang, K. (2021). The application of capsule neural network based CNN for speech emotion recognition. International Conference on Pattern Recognition (ICPR), 9356- 9362.

Yunhao, Z., and Xiaoqing S. (2023). Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error-correcting output codes (ECOC). *Scientific Reports*, 13: 1- 18.