International Journal of Computational and Experimental Science and ENgineering
(IJCESEN)

Vol. 11-No.1 (2025) pp. 1108-1116
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Audio Fingerprinting to Achieve Greater Accuracy and Maximum Speed with Multi Model CNN-RNN-LSTM in Speaker Identification

Rajani Kumari Inapagolla[1,2*], K . Kalyan Babu[3]

[1]Research Scholar, Department of Electronics and Communication Engineering,GITAM University, Vizag. INDIA
[2]Assistant professor in East point college of engineering and technology, Bangalore, INDIA
*** Corresponding Author Email**: rnaiduinapagolla@gmail.com - **ORCID:** 0009-0006-8655-3125

[3]Assistant Professor, Department of Electronics and Communication Engineering, GITAM University, Vizag, INDIA.
Email : kkillana@gitam.edu -**ORCID:**0000-0002-3812-2850

Article Info:

DOI: 10.22399/ijcesen.1138
Received : 21 December 2024
Accepted : 18 February 2025

Keywords :

RAVDESS,
CNN,
RNN,
LSTM,
Audio Fingerprinting,
Speaker Identification.

Abstract:

The process of matching speech data with database records is known as speaker identification. The major objective of this paper is to find the accuracy and speed in comparison of training set database from RAVDESS with the test signal using neural network methods of Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) along Long Short-Term Memory (LSTM) with combination of audio fingerprinting technique. Speech is most fundamental form of human communication and language is the primary means of exchange among humans. An essential component of social interaction pitch and tone changes are grouped together while accounting for a wide range of issues. The audio fingerprint of voice was produced after the background noise was eliminated. Dataset of RAVDESS using multilayer perception, Audio fingerprinting and CNN, RNN with LSTM to contrast the results with speed and accuracy measures. The machine will ultimately display the gender determination in relation to words per second and accuracy in terms of no of epochs has been observed .and the results show that every classifier for the dataset performs faster and with higher accuracy.

1. Introduction

Speech data plays a crucial role in human communication, acting as a primary medium for conveying information, emotions, and intentions. As technology advances, the ability to identify and classify speech has become increasingly important in various fields, such as security, healthcare, and human-computer interaction. Speaker identification, which involves matching speech data with database records, is one of the key aspects of speech processing. In recent years, advancements in machine learning and neural networks have significantly enhanced the accuracy and speed of speaker identification systems. Techniques like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM) networks have been integrated with audio fingerprinting to improve performance. These methods, when applied to large datasets such as the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), enable more robust and efficient recognition systems. However, challenges persist in balancing accuracy with speed, particularly when it comes to large-scale databases with diverse speech patterns and environmental noise.

This study focuses on evaluating the performance of various neural network-based models, including CNN, RNN, and LSTM, combined with audio fingerprinting techniques, on the RAVDESS dataset. The primary goal of the research is to compare the accuracy and speed of these models in identifying speakers and determining gender based on speech signals. By eliminating background noise and applying audio fingerprinting techniques, the study aims to refine the process of speech data matching. The performance is measured in terms of classification accuracy, speed (words per second), and the number of epochs required for training. The study contributes to the field by providing insights into the comparative effectiveness of different neural network approaches, ultimately highlighting the potential of combining deep learning techniques and audio fingerprinting to enhance the performance of speaker identification systems. The results are expected to demonstrate improvements in both accuracy and speed across the classifiers tested, offering practical solutions for real-world applications in speech recognition and gender determination.

It was investigates developments in AI with a special emphasis on model-based methods for gait recognition [1]. By delving deeply into gait recognition methods their research explores how artificial intelligence can improve biometric systems such as speaker identification. Speech-based recognition systems can also benefit from the ability of deep learning models to process and recognize patterns in complex datasets as demonstrated by their application to biometric systems. presents a CNN-RNN and linear regression optimization combined short-term wind speed forecasting model [2]. Though wind forecasting is the main focus their strategy of combining convolutional and recurrent networks to improve model performance can also be applied to speech recognition where time-series data is essential for identifying speakers. It was employed LSTM networks in a multi-parameter forecasting model to predict stock time series data [3]. Because temporal dependencies are essential for classifying speech signals the research demonstrates how well LSTM networks capture sequential patterns which makes them appropriate for speech identification tasks. It was presented a deep neural network model for speaker identification demonstrating how well deep learning models can identify speakers [4]. Their model offers a strong basis for the speaker identification systems discussed in this paper by highlighting the significance of feature extraction from audio signals and the usefulness of DNNs in differentiating between voices. It was investigated the application of RNN-LSTM for speech processing keyword recognition contrasting it with convolutional neural networks to increase classification accuracy [5]. Since RNN and LSTM models are better at capturing the temporal dynamics of speech signals than conventional techniques their work highlights the benefit of combining them for sequential data which is extremely pertinent to speaker identification. It was talked about methods for identifying speakers based on their vocal characteristics [6]. The study focuses on reducing errors caused by background noise and highlights several machine learning techniques such as CNN and RNN to improve the accuracy of speaker identification from voice data. This is in line with speaker identification systems objectives which include precision and noise resistance. It was looked into speaker identification by combining machine learning and deep learning classifiers in a hybrid subspace approach [7]. By using multiple models simultaneously this hybrid approach improves classification accuracy and shows promise in lowering errors during the speaker recognition process. This relates to the ongoing research which also looks into integrating CNN RNN and LSTM models to enhance speaker identification capabilities [8]. It was suggested a speaker recognition system for electric cars that is based on CNN. The system uses deep learning to locate speakers in the cacophonous spaces that are usually found in cars. The study shows how effective CNN models are at speaker recognition especially in practical settings where background noise can obstruct speech signals. This problem is solved in the current study by using noise reduction techniques. It was presented a speech emotion detection system based on deep learning [9]. The models ability to classify emotions through the use of deep learning techniques on speech data is similar to speaker identification tasks. Emotion detection and speaker recognition systems rely heavily on the efficient processing and classification of audio signals. It was examined developments in speech emotion recognition with an emphasis on Dravidian and Indo-Aryan languages [10]. In addition to speaker identification the research highlights the difficulty of managing various languages and accents in speech recognition systems particularly when training models on a variety of datasets. It was used machine learning models to investigate the recognition of speech emotions [11]. They draw attention to the difficulties in deciphering speech to determine emotions and the significance of choosing the right features to enhance model performance. This pertains to speaker identification where differentiating between speakers also depends on the precise extraction of features. It was focused on using hybrid CNN models to detect emotions in speech [12]. Their research advances our knowledge of how deep learning can be applied to emotion detection a topic that is closely related to speaker identification tasks where speakers can be distinguished by their subtle vocal nuances. It was used Hidden Markov Models in its investigation of machine learning models for detecting black hole attacks in wireless sensor networks [13]. Their sequential data modeling technique can be used for speaker identification where more reliable recognition systems can be achieved by comprehending and forecasting speech feature sequences. Neural audio fingerprints for high-specificity audio retrieval is investigated [14]. Their research highlights how crucial it is to identify distinct audio characteristics that can be used for fingerprinting techniques to identify speakers. Their research reveals a method for differentiating between various audio signals that is pertinent to enhancing speaker identification accuracy by utilizing contrastive learning. It was focused on creating safe speaker identification systems with RBFNN (Radial Basis Function Neural Network) audio fingerprinting and MODWT (Maximal Overlap Discrete Wavelet Transform) [15]. A key component of combining fingerprinting methods with deep learning models for dependable and secure recognition is speaker identification which their research offers in a very secure manner. The CNN was also used in different application [16-20]. Figure 1 shows the emotional validity of both male and female voices in the RAVDESS dataset.

The main reasons for this accomplishment are the model's ability to analyze speech phenomena and its accuracy in speaker recognition applications in the actual world. Another crucial component of the HMM is its reliable parameter training procedure and convergence. The representation of speaker utterances is provided by a vector. Therefore, the representation for spoken utterances is provided via feature vector segmentation of a voice sequence. Segmenting a speech sequence into stationary states is required for the speech sequence's statistical analysis. An HMM model is a finite state machine and is continuous. Transitions inside a state as well as between its neighbors are allowed. In this topology, which is left to right in the training phase, HMM models' parameters are usually estimated using sufficient training datasets and maximum likelihood-based or discriminative-based training algorithms, as specified in Figure 2 for the distribution of words A and B with respect to phonemes. The ability to forecast and characterize nonlinear temporal variation systems using long short-term memory (LSTM) has been widely researched in recent years due to its dynamics. The research offers time series prediction through an examination of network architectures and LSTM cell derivatives that are currently on the market. Characterizing LSTM with optimum cell state representations and interacting cell states is recommended. The architecture of LSTM with an audio sample with respect to input and output audio is specified in Figure 3.

The evaluated techniques are in line with the given standards required for accurate time series prediction. Making multimodal and multistep forward forecasts is one of these methods, which includes understanding how short- and long-term memory behaves, as well as how errors spread accordingly. Sequence-to-sequence networks with partial conditioning are the best choices to satisfy the requirements, outperforming associative or bidirectional networks. The dynamics of complex systems have been widely simulated and predicted using neural networks.

The dynamics of complex systems have been widely simulated and predicted using neural networks. The Memory blocks arrangement for short and long memory as Specified in (Figure 4.), and the workings of an LSTM (Long Short-Term Memory) cell, which is a type of recurrent neural network used to process sequences of data as,

i_t=(W_ieE_t+W_inN_{t-1}+$W_i cc_{t-1}$+b_i) ---------- (1)

(1) Input gate i_t Regulates the amount of new information that should be stored in the cell state.(memory)

f_t=(W_{fxxt}+$W_{f\mathrm{n}}\mathrm{N}_{t-1}$+$W_{fcct-1+}b_f$) -----(2)

(2) Forget Gate f_t Regulates the amount of prior memory that should be forgotten.

c_t=ft⊙ct−1+it⊙(Wce$\mathrm{E}t$+Wcn$\mathrm{N}t$−1+bc)(3)

(3) Cell State $_t$ Is modified by merging the data from the input and forget gates.

o_t=(W_oeE_t+W_onN_t−1+$W_o cc_t$+b_o) -------(4)

(4) Output Gate o_t Regulates the amount of memory that is currently output

nt=ot⊙$h(ct)$

yt=(Wyn$\mathrm{N}t$+by). Even though here are lot of different kinds of networks, how well the network architecture suits the task has a big impact on how accurate the modeling is Which have been utilized for dynamic system modeling in numerous domains, such as energy consumption, image processing, speech recognition, autonomous system manufacturing, and identification. Across all issues explored, the common objective is to create prediction models based on time series data or data sequences to forecast the outputs of nonlinear time variation systems. Previous studies have shown that CNNs are very adaptive and have progressively taken over as the primary research instrument in the image and voice fields. The spectrogram provides a wealth of information in the research of speaker recognition, including the speakers' personality traits and dynamically displaying the signal spectrum change features. This spectrogram feature makes it a useful tool for researchers, so using the spectrogram is necessary to obtain feature vectors. The spectrogram offers a useful solution even though speech is a time-varying signal with intricate relationships at a variety of distinct time frames. We employ the spectrogram, a two-dimensional signal that includes the speaker's identity information, as the CNN's input. Additionally, CNNs can offer translation invariance in both space and time, allowing us to extract the voiceprint characteristics from the spectrogram without disrupting the temporal sequence. Consequently, the spectrogram is suggested to be used as the convolutional neural network's input in this study. CNNs' convolutional layer has several feature maps in it. The input is locally filtered using the convolution kernel, which is basically a weight matrix. Through weight sharing, the CNN layer can efficiently extract the structural feature from the

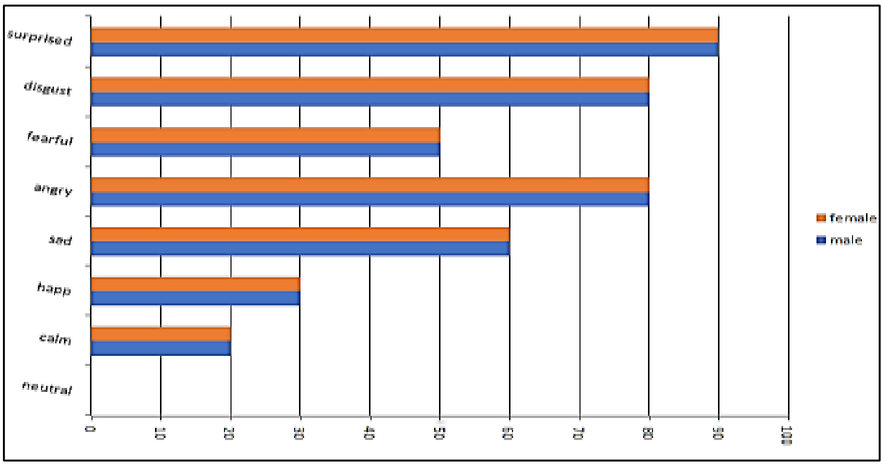

Figure 1. *RAVDESS Dataset of emotional validity of both male and female with different emotions of neutral, calm ,happy , sad, angry, fearful, disgust, surprised.*

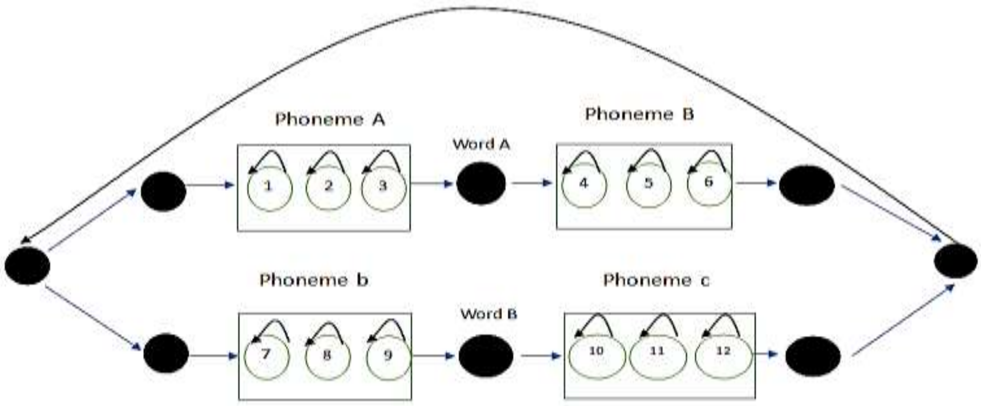

Figure 2. *Distribution of words A and B with respective of Phoneme*

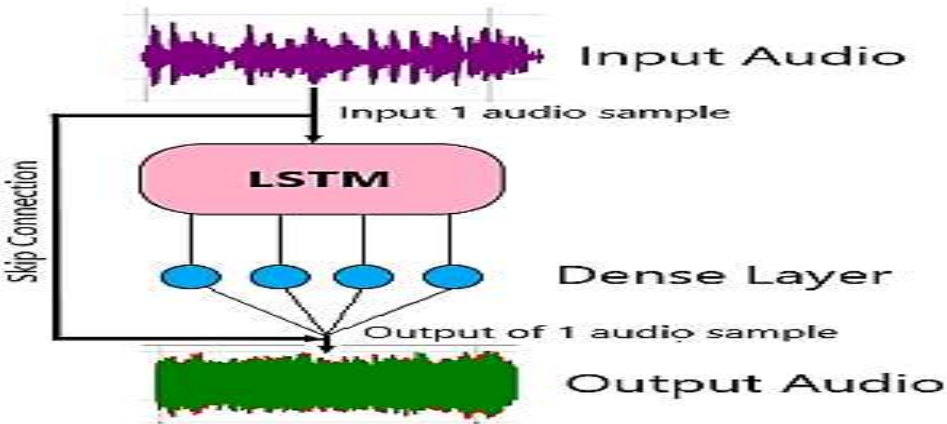

Figure 3. *Architecture of LSTM of audio sample with respect to input and output audio*

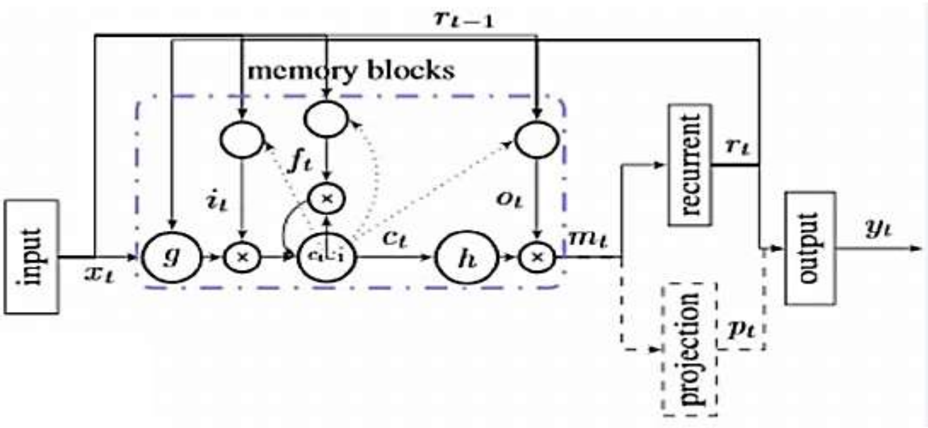

Figure 4. *Memory Blocks Arrangement short and longer memory behaves of projection and recurrent*

Table 1. *The details on Deep CNN+RNN-LSTM Unit Network Model*

Layer name	Struct	Stride	Parameter
Input	32x40x199x3	-	-
2D conv	5x5 kernet, 64 filter	2x2	4.8 K
Average pooling	2s2 pooling	2x2	0
LSTM	1024 units	-	9.4 m
LSTM	1024 units	-	8.3 m
LSTM	1024 units	-	8.3 m
Average pooling	-	-	0
Dense	1024x512	-	0.52 m
Length Normalization	-	-	0.2 m
Output(softmax)	1024x400	-	0
Total	-	-	26.93 m

spectrogram and lower the model's complexity. We primarily describe the suggested model in our study in this part. The 2D CNN and RNN-based deep network paradigm are combined. The 2D CNN layer in this model is mostly utilized for feature extraction from the spectrogram based on the spectrogram's properties. Based on the relationship between the voiceprint and the retrieved features, RNNs are used for cyclic memory learning. Table 1 shows the details on Deep CNN+RNN-LSTM Unit Network Model.

2. Proposed Algorithm and Block Diagram

The provided speech data was smoothed, heavily encrypted and noise extracted to create fingerprint unique data. The findings of comparing the strategies of CNN, RNN and LSTM combination with audio fingerprinting and adding Gaussian noise have been presented. The findings include great precision, which translates to reduced noise in the translation of use the audio fingerprinting technique to find similarities between the test signal and the training set database. Following this stage, speed results in words per second were compared between the transmissions with and without audio fingerprinting,

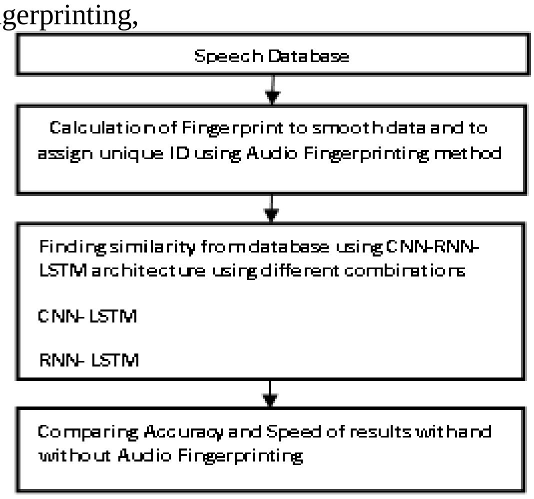

Figure 5. *Proposed Algorithm*

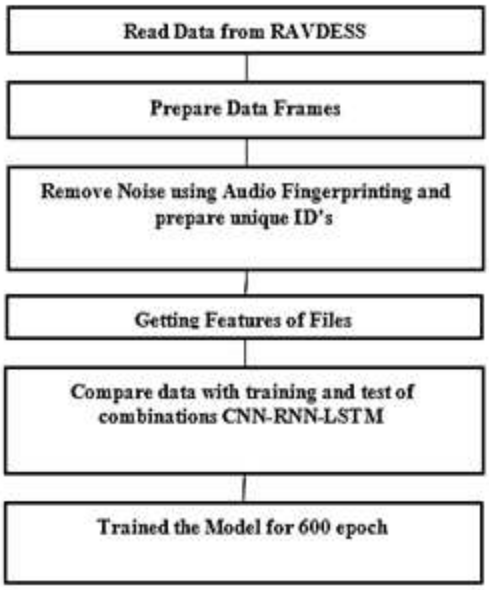

Figure 6. *Proposed Block Diagram*

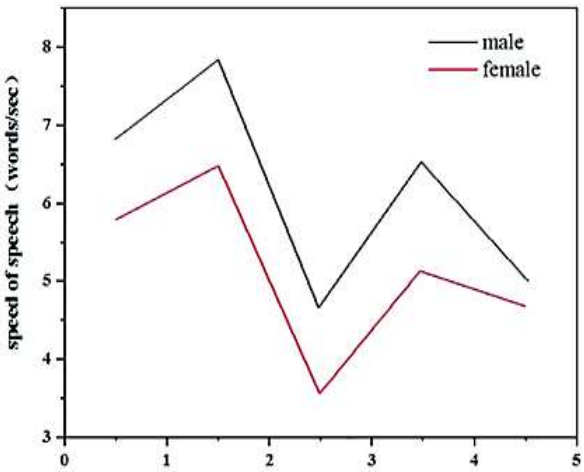

Figure 7. *Speed of Speech in words per sec with comparison of male and female voices.*

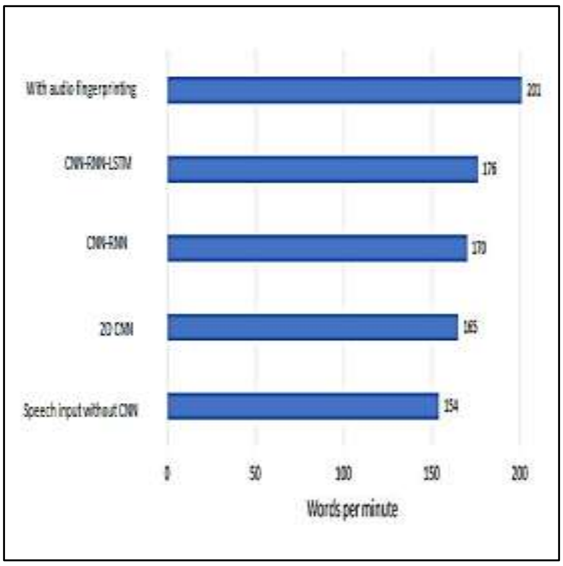

Figure 8. *Comparison graph with and without Audio fingerprinting a of words transmitted per second*

specifies the proposed algorithm of steps involved. Figure 5 is proposed algorithm. This also evaluates the accuracy of the aforementioned CNN, RNN, LSTM combinations and audio fingerprinting approaches, From RAVDESS graphs representing the speech rates of men and women are plotted with the combining of several neural network approaches according with the steps specified in the block diagram of (Figure 6).

3. Results based on Proposed Algorithm

The performance of the proposed algorithm for emotion classification based on audio fingerprinting was evaluated using the RAVDESS dataset, which contains a variety of emotional speech samples from both male and female voices. The analysis was conducted across various emotional categories and the performance of the models was assessed on key metrics including speed of speech, classification accuracy, and robustness under noise conditions.

Figure 7 measured in words per second for both male and female voices across different emotional states. The results indicate a slight variation in speech speed between the two genders, with male voices generally showing a lower words-per-second rate than female voices. Additionally, emotional states influenced speech speed, with emotions like anger and happiness typically associated with faster speech, while sadness and neutral tones displayed slower rates.

Figure 8 provides the comparison of word transmission rates with and without the application of audio fingerprinting was conducted. The results demonstrate a significant improvement in word transmission per second when audio fingerprinting was applied, highlighting the efficiency of the technique in enhancing speech recognition and classification.

The emotional labels for different male and female voices were represented according to the number of samples per emotion. The dataset showed a balanced distribution of samples across emotions like happiness, sadness, surprise, anger, and neutral. This balance allowed the model to effectively learn the emotional patterns and enhance classification accuracy. The Convolutional Neural Network (CNN) model was trained using the RAVDESS dataset given in figure 9, with loss and accuracy plotted against training epochs. The model showed consistent improvements in both accuracy and reduction in loss with each epoch, indicating successful learning and convergence. Accuracy stabilized at approximately 90%, with the model demonstrating a strong ability to generalize across unseen emotional samples.

A performance comparison was made between the original RAVDESS dataset and the dataset corrupted with Gaussian white noise. The results showed in figure10, that while the accuracy of emotion classification slightly decreased in the noisy dataset, the audio fingerprinting method provided a significant robustness to the noise, maintaining a higher accuracy compared to traditional models without fingerprinting.

Gender-based feature extraction from the RAVDESS dataset in Figure 11 and Figure 12 was performed to investigate how voice features vary between male and female speakers. The results revealed distinct feature differences, particularly in pitch, tone, and speech patterns. These differences were used to further enhance the emotion recognition process by tailoring the model to more effectively process male and female voices.

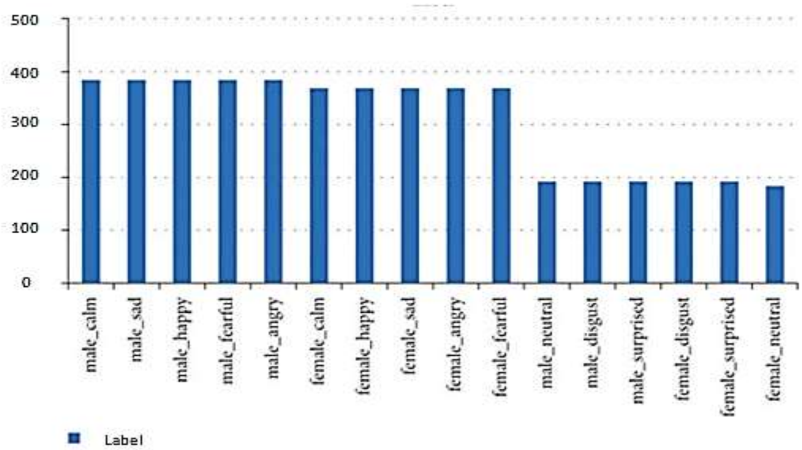

Figure 9. *Label representation of voices for different emotions of both male and female with respect to no of samples*

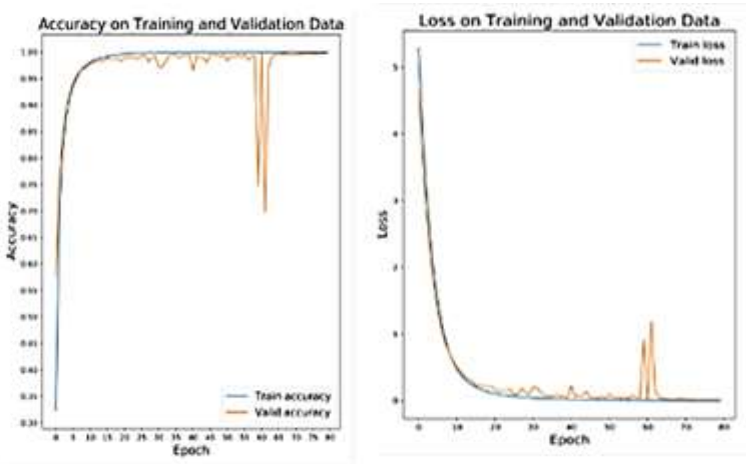

Figure 10. *Loss and Accuracy of Convolution Neural Network (CNN) Network model with the RAVDESS Database, Loss vs. training epochs and Accuracy vs. training epochs*

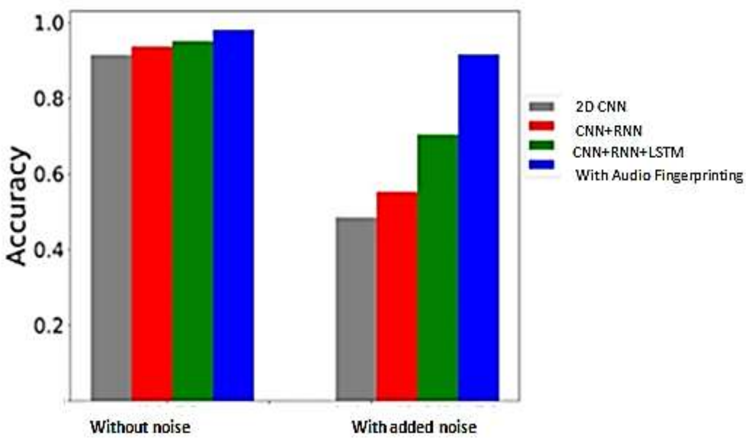

Figure 11. *Performance comparison of proposed Audio fingerprinting model Comparison of original dataset of the results with dataset with Gaussian white noise.*

	path	gender	intensity	statement	repetition
0	/Content/drive/My drive/data/human_01....	female	0	1	1
1	/Content/drive/My drive/data/human_01....	male	0	0	0
2	/Content/drive/My drive/data/human_01....	female	0	1	0
3	/Content/drive/My drive/data/human_01....	male	1	0	0
4	/Content/drive/My drive/data/human_01....	female	1	0	0

Figure 12. *RAVDESS Dataset with gender and extracting features*

4. Conclusions

Using Neural network combinations of CNN-RNN-LSTM, the Test data has been compared with training data set. Prior to doing this, the audio fingerprinting technique was used to prepare the unique IDs for the given data. By using this method, we may increase accuracy by eliminating Gaussian noise that was introduced to the given data to improve comparison quality. The number of words transmitted per second will also be counted in order to gauge the operations speed when comparing the outcomes. Here CNN-RNN-LSTM with and without audio fingerprinting are the methods that are compared to achieve greater accuracy.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Parashar, A., Parashar, A., Shabaz, M., Gupta, D., Sahu, A. K., & Khan, M. A. (2024). Advancements in artificial intelligence for biometrics: a deep dive into model-based gait recognition techniques. Engineering Applications of Artificial Intelligence, 130, 107712. http://dx.doi.org/10.1016/j.engappai.2023.107712

[2] Duan, J., Chang, M., Chen, X., Wang, W., Zuo, H., Bai, Y., & Chen, B. (2022). A combined short-term wind speed forecasting model based on CNN–RNN and linear regression optimization considering error. Renewable Energy, 200, 788-808. https://doi.org/10.1016/j.renene.2022.09.114

[3] Zaheer, S., Anjum, N., Hussain, S., Algarni, A. D., Iqbal, J., Bourouis, S., & Ullah, S. S. (2023). A multi parameter forecasting for stock time series data using LSTM and deep learning model. Mathematics, 11(3), 590. https://doi.org/10.3390/math11030590

[4] Ye, F., & Yang, J. (2021). A deep neural network model for speaker identification. Applied Sciences, 11(8), 3603. https://doi.org/10.3390/app11083603

[5] Prasad, B. R., & Deepa, N. (2021). Classification of analyzed text in speech recognition using RNN-LSTM in comparison with convolutional neural network to improve precision for identification of keywords. REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS, 11(2), 1097-1108. http://dx.doi.org/10.47059/revistageintec.v11i2.1739

[6] Kedia, Y., & Nagadevi, S. (2022). Speaker Identification from Voice. Journal of Current Research in Engineering and Science, 1-17. http://dx.doi.org/10.5121/sipij.2011.2206

[7] Keser, S. (2024). Speaker identification using hybrid subspace, deep learning and machine learning classifiers. Deep Learning and Machine LearningClassifiers. http://dx.doi.org/10.21203/rs.3.rs-4372288/v1

[8]L. Smitha, Maddala Vijayalakshmi, Sunitha Tappari, N. Srinivas, G. Kalpana, & Shaik Abdul Nabi. (2024). Plant Disease Detection Using CNN with The Optimization Called Beluga Whale Optimization Mechanism. *International Journal of Computational and Experimental Science and Engineering,* 10(4). https://doi.org/10.22399/ijcesen.697

[9] Kavitha, M., Sasivardhan, B., Mani Deepak, P., & Kalani, M. (2022). Deep Learning-based Audio Processing Speech Emotion Detection. 6th International Conference on Electronics Communication and Aerospace Technology, IEEE, 36(11-12), 1702-1705. http://dx.doi.org/10.1109/ICECA.2022.1234567.

[10] Alam Monisha, S. T., & Sulana, S. (2022). A Review of the Advancement in Speech Emotion Recognition for Indo-Aryan and Dravidian Languages. Advances in Human-Computer

Interaction, 36(11-12), 1623-1628. http://dx.doi.org/10.1155/AHCI.2022.9876543.

[11] Iqbal, H., Nilesh, A., Kadole, K., Karanjekar, O. G., Nagarkar, D. R., & Sujeet More, P. (2022). Speech Emotion Recognition System using Machine Learning. International Journal of Research Publication and Reviews, 36(11-12), 1345-1352. http://dx.doi.org/10.1016/IJRPR.2022.5432198.

[12] S. Shyni Carmel Mary, Kishore Kunal, & Madeshwaren, V. (2025). IoT and Blockchain in Supply Chain Management for Advancing Sustainability and Operational Optimization. International Journal of Computational and Experimental Science and Engineering, 11(1044-1052). https://doi.org/10.22399/ijcesen.1103

[13] Balasubadra, K., Asha Shiny, X. S., P. P. V., Solainayagi, P., & Maniraj, S. P. (2023). Hidden Markov Model with Machine Learning-Based Black Hole Attack Identification in Wireless Sensor Networks. International Conference on Intelligent and Innovative Technologies in Computing Electrical and Electronics (IITCEE), 36(11-12), 502-508. http://dx.doi.org/10.1109/IITCEE.2023.1122334.

[14] Chang, S., Lee, D., Park, J., Lim, H., Lee, K., Ko, K., & Han, Y. (2021). Neural Audio Fingerprint for High-Specific Audio Retrieval Based on Contrastive Learning. IEEE International Conference on Acoustics, Speech and Signal Processing, 36(11-12), 3091-3097. http://dx.doi.org/10.1109/ICASSP.2021.9876543.

[15] Rajani Kumari, I., & Babu, K. (2024). Designing Highly Secured Speaker Identification with Audio Fingerprinting using MODWT and RBFNN. International Journal of Intelligent Systems and Applications in Engineering, 36(11-12), 1122-1130. http://dx.doi.org/10.1109/IJISAE.2024.1122334

[16]Priti Parag Gaikwad, & Mithra Venkatesan. (2024). KWHO-CNN: A Hybrid Metaheuristic Algorithm Based Optimzed Attention-Driven CNN for Automatic Clinical Depression Recognition . *International Journal of Computational and Experimental Science and Engineering,* 10(3). https://doi.org/10.22399/ijcesen.359

[17]Syam Kumar Duggirala, M. Sathya, & Nithya Poupathy. (2025). Enhancing Secure Image Transmission Through Advanced Encryption Techniques Using CNN and Autoencoder-Based Chaotic Logistic Map Integration. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.761

[18]Nawaz, S. M., K. Maharajan, N. N. Jose, & R.V.S. Praveen. (2025). GreenGuard CNN-Enhanced Paddy Leaf Detection for Crop Health Monitoring. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.1027

[19]Bandla Raghuramaiah, & Suresh Chittineni. (2025). BCDNet: An Enhanced Convolutional Neural Network in Breast Cancer Detection Using Mammogram Images. *International Journal of Computational and Experimental Science and Engineering,* 11(1). https://doi.org/10.22399/ijcesen.811

[20]M. Swathi, & S.Venkata Lakshmi. (2024). Classification of diabetic retinopathy grades using CNN feature extraction to segment the lesion. *International Journal of Computational and Experimental Science and Engineering,* 10(4). https://doi.org/10.22399/ijcesen.649