STM JOURNALS

Journal of
Communication Engineering & Systems

ISSN: 2249-8613 (Online)
ISSN: 2321-5151 (Print)
Volume 15, Issue 1, 2025
January–April
DOI (Journal): 10.37591/JoCES

https://journals.stmjournals.com/joces

Review JoCES

Comparative Analysis of MCNN and RCNN for Speech Emotion Recognition Using Gender Information

Gowtami Annapurna Dinavahi[1,*], Chandini R.[2], Swetha Akshaya P.[2], Kavya D.[2], Vandana M.[2]

Abstract

Speech emotion recognition is a speech processing task and a computer-based approach designed to identify and classify the emotions conveyed in audio signals. The aim of this system is to evaluate a speaker's emotional state, such as happiness, anger, sadness, or frustration, by analyzing their speech patterns, which include prosodic features like pitch, frequency, and rhythm. Speech emotion recognition is used in various real-life scenarios that include Customer Service, Healthcare, Education, Human-Computer Interactions, Market Research, etc. This study presents a comparative analysis of Mixed Convolutional Neural Networks (MCNN) and Residual Convolutional Neural Networks (RCNN) for speech emotion recognition, specifically considering the incorporation of gender information. Mixed Convolutional Neural Networks (MCNN) and Residual Convolutional Neural Networks (RCNN) models are trained and evaluated on datasets consisting of speech samples labelled with different emotions, along with gender information. At the end, the above models are evaluated on following datasets: SAVEE, RAVDEES, EMODB. This comparative analysis will aid in understanding the strengths and limitations of each model and guide researchers in selecting the most suitable model for speech emotion recognition considering gender information.

Keywords: Customer segmentation, product segmentation, clustering, business, clients

INTRODUCTION

Speech emotion recognition simulates the human process of perceiving and understanding emotions. It extracts prosodic features from collected WAV files and determines the relationship between these features and human emotions. This technology is widely utilized in human-computer interaction. In the educational context, analyzing online learners' emotional states can enhance teaching quality and improve the learning experience. Emotion recognition from speech holds potential in various real-life applications such as customer service, healthcare, education, human-computer interactions, and market research. The ability to accurately discern emotions from speech not only enhances the efficiency and effectiveness of communication systems but also opens doors to personalized and empathetic interactions. One crucial aspect that adds complexity to the task of speech emotion recognition is the consideration of gender information. Gender is a multifaceted construct that influences various aspects of speech production, including pitch, frequency, intonation, and vocal quality. As such, incorporating gender information into emotion recognition models can significantly enhance their performance and robustness across diverse demographic groups. This project focuses on exploring the role of gender information in speech emotion recognition through a comparative analysis of two prominent neural network architectures: Mixed Convolutional Neural

***Author for Correspondence**
Gowtami Annapurna Dinavahi
E-mail: dinavahigowtami@gvpcew.ac.in

[1]Assistant Professor, Department of Computer Science Engineering, Gayatri Vidya Parishad College of Engineering for Women, Visakhapatnam, Andhra Pradesh, India
[2]Student, Department of Computer Science Engineering, Gayatri Vidya Parishad College of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

Received Date: October 25, 2024
Accepted Date: November 07, 2024
Published Date: December 31, 2024

Citation: Gowtami Annapurna Dinavahi, Chandini R., Swetha Akshaya P., Kavya D., Vandana M. Comparative Analysis of MCNN and RCNN for Speech Emotion Recognition Using Gender Information. Journal of Communication Engineering & Systems. 2025; 15(1): 1–10p.

Networks (MCNN) and Residual Convolutional Neural Networks (RCNN). By training and evaluating these models on datasets annotated with both emotional states and gender information understanding how gender influences the expression and perception of emotions in speech is not only essential for advancing the field of speech emotion recognition but also holds implications for designing more inclusive and context-aware communication systems.

RELATED WORK

Singh and Prasad [1]: "Speech emotion recognition system using gender dependent convolution neural network": The study aims to enhance SER accuracy by considering gender-specific features in speech. They preprocess speech data to extract pitch, intensity, and formant frequencies, along with gender-related features. The gender-dependent CNN architecture includes separate processing paths for male and female speech data to capture gender-specific acoustic traits. The model is trained and evaluated on a labelled dataset, showing superior performance compared to traditional CNN models in recognizing emotions from speech. Integrating gender-specific features greatly enhances the accuracy of speech emotion recognition (SER), underscoring the significance of accounting for gender influences in speech emotion analysis.

Zhang et al. [2]: "A Deep Learning Method Using Gender-Specific Features for Emotion Recognition": The study aims to improve emotion recognition accuracy by considering gender-related characteristics in speech. They preprocess the data to extract relevant features and propose a deep learning architecture that incorporates gender-specific information. The model is trained and assessed using a labeled dataset, showcasing better performance than traditional entity recognition methods. By incorporating gender-specific features, the proposed approach achieves enhanced accuracy in recognizing emotions from speech data. This involves utilizing MLP for gender classification, analyzing the impact of various speech emotion features on male and female speech, determining optimal feature sets for recognizing emotions in both genders, and training and testing CNN and Bi-LSTM models using the respective speech emotion feature sets for males and females.

Tigga and Garg [3]: "Speech Emotion Recognition for multiclass classification using Hybrid CNN-LSTM": This is focused on a novel approach to speech emotion recognition (SER) by combining three datasets for emotion recognition, implementing MFCC for feature extraction, and using a hybrid CNN-LSTM technique to recognize emotions in audio signals. This hybrid model combines CNNs for understanding patterns in the speech data and LSTM for remembering long sequences of information. Their main goal is to accurately identify different emotions expressed in speech, which is tricky because there are many emotions to consider. The paper explains how they built and trained this hybrid model, showing how it learns from data to understand emotions in speech.

Fu et al. [4]: "Cross-Corpus Speech Emotion Recognition Based on Multi-Task Learning and Subdomain Adaptation": Fu *et al*. present a novel approach for cross-corpus speech emotion recognition (SER) utilizing Multi-Task Learning (MTL) and Subdomain Adaptation. Their model addresses feature distribution discrepancies across corpora, enhancing emotion recognition performance. By employing a deep denoising auto-encoder (DDAE) as a shared feature extraction network and adding task-specific layers, the model improves feature representation. Additionally, a subdomain adaptation algorithm aligns emotional and gender features between source and target domains, mitigating distribution differences. The study underscores the importance of considering gender factors in emotion recognition and provides insights for addressing cross-corpus SER challenges through MTL and subdomain adaptation.

Wani et al. [5]: "A Comprehensive Review of Speech Emotion Recognition Systems": Speech emotion recognition (SER) is a crucial element in Human-Computer Interaction (HCI) and advanced speech processing systems. It highlights the necessity for more robust algorithms to enhance the performance of SER systems and improve accuracy rates. The paper describes the critical role of speech in human communication, highlighting its richness in conveying both linguistic and paralinguistic information, including emotions. Emotion recognition has been historically focused on facial

expressions but has recently expanded to include speech signals. Speech emotion recognition (SER) has become increasingly popular in the realm of human-computer interaction.

Nicolini and Ntalampiras [6]: Nicolini and Ntalampiras propose a hierarchical approach to multilingual speech emotion recognition (SER), focusing on gender and emotional state prediction. They employ three classifiers: k-NN, transfer learning with YAMNet, and BiLSTM neural networks. Their method considers six languages and the big-six emotions, aiming to generalize patterns for emotion recognition across languages. Results show that gender differentiation improves SER performance. YAMNet and k-NN perform well for gender discrimination, while BiLSTM outperforms other classifiers for emotion recognition, especially capturing temporal dependencies. Gender-dependent classification shows improvement when considering female emotional speech.

Latif et al. [7]: Latif *et al.* introduce a novel approach, "Multi-Task Semi-Supervised Adversarial Autoencoding", to enhance speech emotion recognition (SER). This approach combines multiple tasks and semi-supervised learning with adversarial autoencoding techniques. Their method aims to enhance SER by leveraging both labelled and unlabeled data, which is crucial for training models effectively when labelled data is scarce. By incorporating adversarial autoencoding, the model learns to generate realistic speech representations while simultaneously enhancing feature extraction for emotion recognition. The paper elaborates on the design and implementation of this approach, detailing how it utilizes semi-supervised learning and adversarial training to improve SER performance.

Nwe et al. [8]: The paper "Speech Emotion Recognition using Hidden Markov Models" presents a text-independent approach for classifying emotions in speech by employing short-time log frequency power coefficients (LFPC) along with a discrete hidden Markov model (HMM) as the classifier. The study focuses on categorizing emotions into six types: Anger, Disgust, Fear, Joy, Sadness, and Surprise. It utilizes a database of emotional utterances from 12 speakers for training and testing the proposed system. The results demonstrate that the LFPC feature parameters significantly outperform traditional features, such as linear prediction cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC) accuracy, with potential applications in diverse domains such as human-computer interaction and affective computing.

DATASET COLLECTION

This Online Retail II dataset includes all transactions conducted by a UK-based registered non-store online retailer from December 1, 2009, to December 9, 2011 as shown in Figure 1. The company primarily specializes in selling unique giftware for various occasions, with a significant portion of its customers being wholesalers.

Attribute Information

- *InvoiceNo:* Invoice number.
- *Description:* Product (item) name.
- *Quantity:* The number of each product (item) in each transaction.
- *InvoiceDate*: Invoice date and time.
- *UnitPrice:* Unit price.
- *CustomerID:* Customer number.
- *Country:* Country name.

PROPOSED SYSTEM

The proposed system is a speech emotion recognition (SER) method that utilizes a combination of mixed convolutional neural networks (MCNN) and residual convolutional neural networks (RCNN), incorporating gender information as shown in Figure 2. It consists of two stages: gender recognition and emotion recognition.

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	01-12-2010 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 8:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART	6	01-12-2010 8:26	3.39	17850	United Kingdom
536370	22728	ALARM CLOCK BAKELIKE PINK	24	01-12-2010 8:45	3.75	12583	France
536370	22727	ALARM CLOCK BAKELIKE RED	24	01-12-2010 8:45	3.75	12583	France
536370	22726	ALARM CLOCK BAKELIKE GREEN	12	01-12-2010 8:45	3.75	12583	France
536389	22941	CHRISTMAS LIGHTS 10 REINDEER	6	01-12-2010 10:03	8.5	12431	Australia
536389	21622	VINTAGE UNION JACK CUSHION COVER	8	01-12-2010 10:03	4.95	12431	Australia
536389	21791	VINTAGE HEADS AND TAILS CARD GAME	12	01-12-2010 10:03	1.25	12431	Australia
536389	35004C	SET OF 3 COLOURED FLYING DUCKS	6	01-12-2010 10:03	5.45	12431	Australia
536389	35004G	SET OF 3 GOLD FLYING DUCKS	4	01-12-2010 10:03	6.35	12431	Australia
536390	22962	JAM JAR WITH PINK LID	48	01-12-2010 10:19	0.72	17511	United Kingdom
536390	22963	JAM JAR WITH GREEN LID	48	01-12-2010 10:19	0.72	17511	United Kingdom

Figure 1. Dataset.

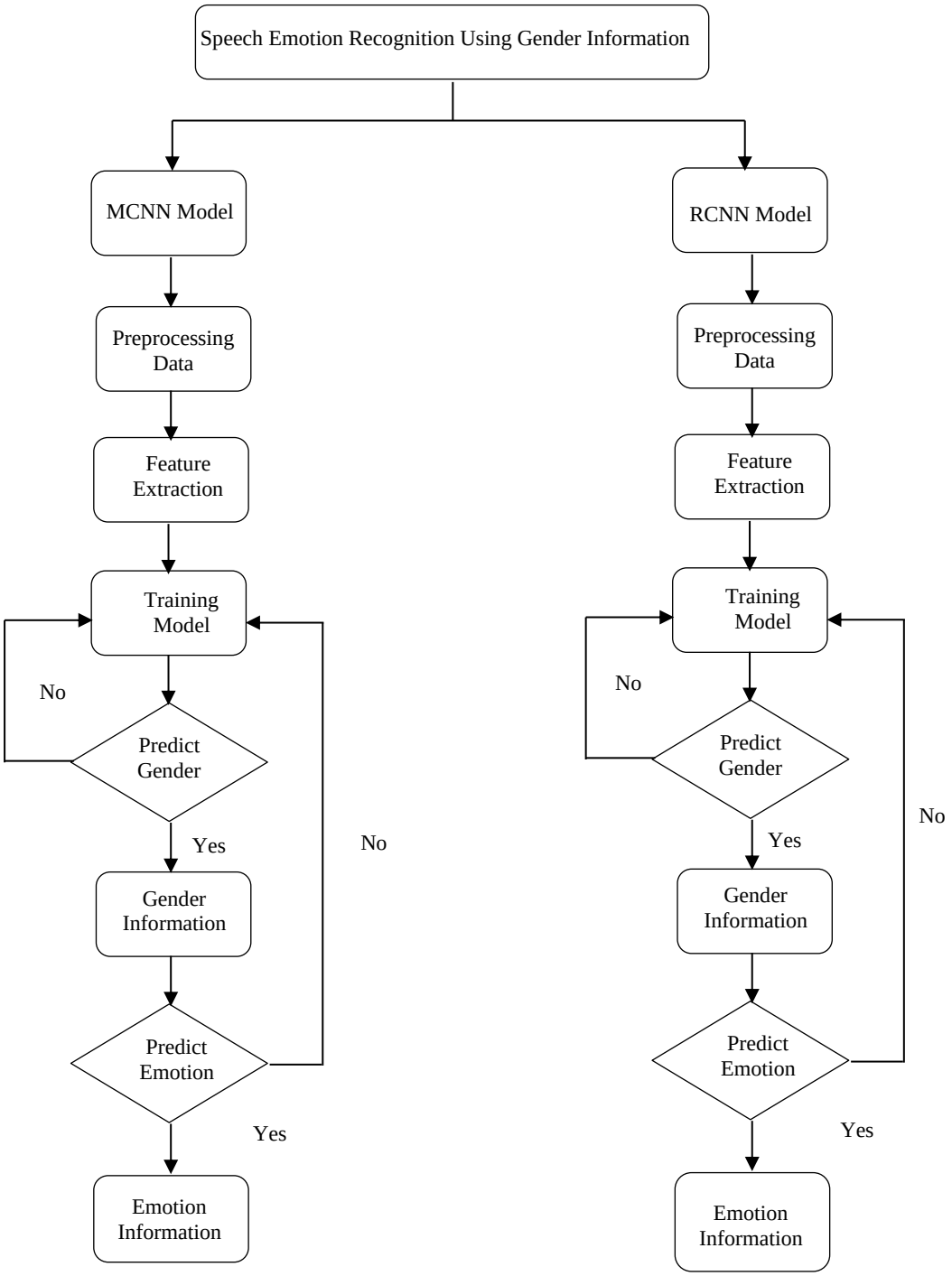

Figure 2. Proposed system.

In this project, mel-frequency cepstral coefficients (MFCC) were utilized as features to classify speech data into various emotion categories using convolutional neural networks (Mixed CNN and Residual CNN). The implementation of these neural networks allows for the classification of multiple emotion types within variable-length audio signals in real-time environments.

METHODOLOGY

The Figure 3 describes the flow of the project starting with the collection of data from different datasets. This data is then pre-processed, gender features are extracted and emotions are labelled with some numbers.

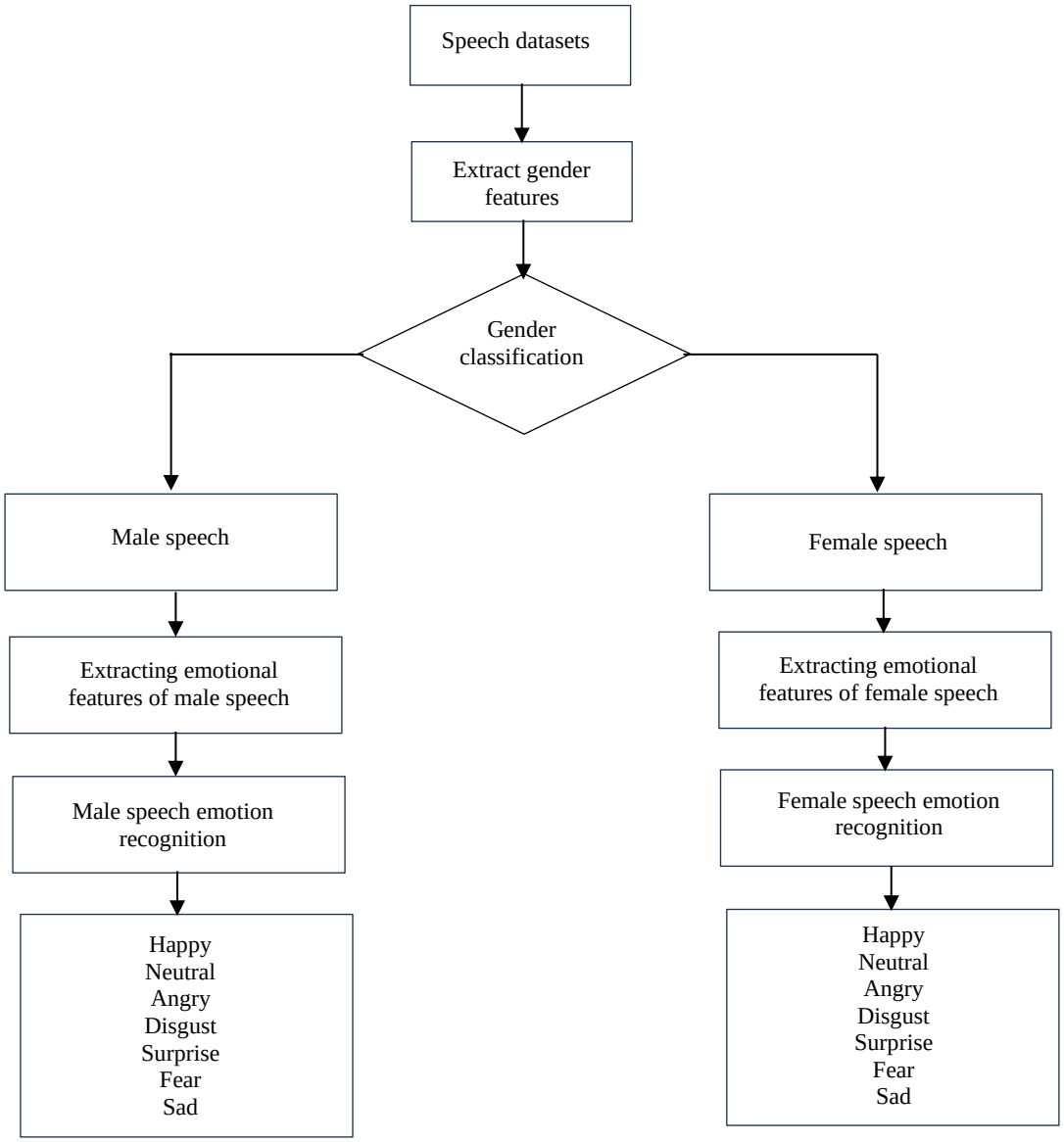

Figure 3. Architecture diagram.

Algorithms Used

Deep Learning Algorithm

Convolutional Neural Network

A Convolutional Neural Network (CNN) is a specific type of artificial neural network employed for various applications, including image recognition, object detection, and speech processing. According to this project, aim is to influence both the prosodic features of speech signals and the contextual information provided by gender to improve emotion classification accuracy by including gender information into the network architecture. Convolutional layers are responsible for extracting hierarchical features from the input spectrograms.

Transformer Models

Mixed Convolution Neural Network

Mixed CNN involves combining different types of convolutional layers and architectures within a single network. Mixed CNN are highly customizable and can be tailored to specific tasks or datasets by selecting and combining different convolutional layers. Features are analyzed at various scales and levels of abstraction, incorporating Coordinate Attention and A-GRUs structures [9].

Coordinate Attention

Coordinate attention assigns different weights to different parts of input based on their importance, enhancing the focus on key parts. CA is introduced as an alternative to Convolutional Block Attention Module (CBAM) to address limitations in capturing long-term dependencies and spatial range. For calculation, CA involves two main steps: One is coordinate information embedding and coordinate attention generation, integrating channel, and other is spatial coordinate information to generate attention maps. CA efficiently captures spatial range without significant computational overhead, using parallel one-dimensional feature coding. CA is described as a plug-and-play model, flexible, and lightweight, addressing channel and spatial considerations while solving long-term dependency issues.

A-GRUs Structure

Gate Recurrent Unit utilized in the A-GRUs model, represents a special form of Recurrent Neural Network. Each GRU unit possesses the capability to learn time context information, facilitating information propagation through hidden states. GRU features only one hidden state, simplifying its structure and reducing computational complexity. This simplicity allows GRU to converge faster, making it well-suited for SER tasks. This model includes an attention mechanism to weigh input sequence information and discern significant features. The coordination attention module improves the CA-MCNN architecture by adding position information to channel attention.

Residual Convolution Neural Network

Residual CNN utilizes residual learning to address the vanishing gradient problem in deep neural networks. This is accomplished by incorporating skip connections, which enable the network to learn residual functions. These connections aid in training deeper architectures by creating shortcut paths for gradient flow. The skip connections in ResCNNs help in mitigating the degradation problem, allowing for more stable and efficient training of very deep networks. While ResCNNs introduce skip connections, the basic architecture remains simpler compared to mixed CNNs, which may involve more diverse layers and architectures. Traditional speech emotion recognition involves two steps: feature extraction from speech signals and classification using a classifier [10].

RESULTS AND ANALYSIS

On clicking Models button, the user can select from the two models which they want to use as shown in Figures 4 and 5. It will then direct the user to the page where they can upload an audio file and obtain the emotion.

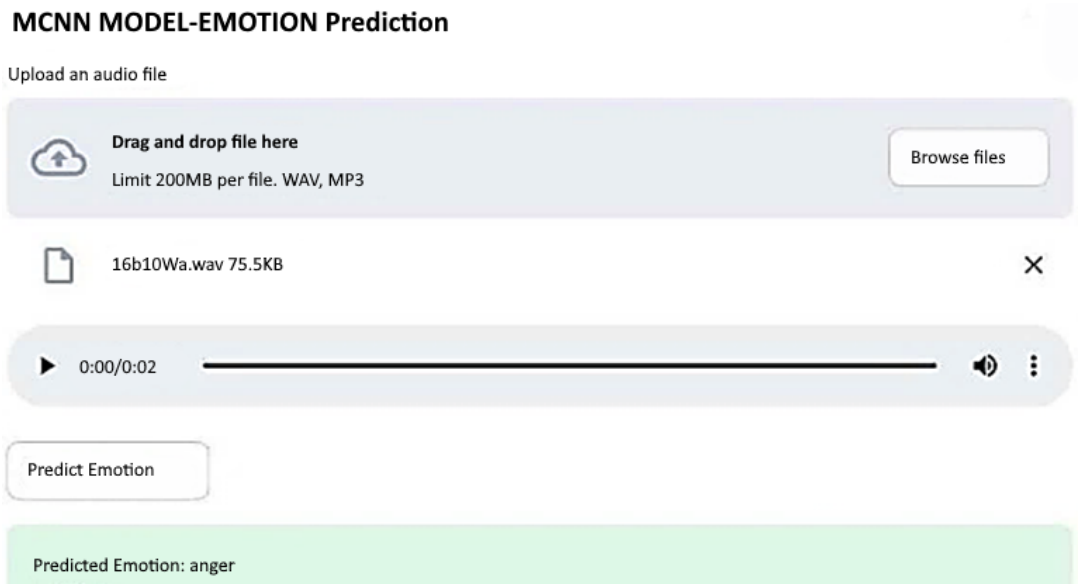

Figure 4. MCNN model.

The below graph consists of 'Actual' and 'Predicted' emotion values, with each line representing an emotion and its corresponding prediction. Then, it counts the occurrences of each emotion in both the 'Actual' and 'Predicted' columns. The blue line plot represents the counts of emotions in the 'Actual' column, and the orange line plot represents the counts of emotions in the 'Predicted' column. Each line on the charts corresponds to an emotion, and its height represents the number of occurrences of that emotion. The x-axis shows the different emotions, while the y-axis shows the count of each emotion. Finally, it adjusts the layout and displays the plots as shown in Figure 6.

Figure 7 calculates the accuracy of predicted emotions compared to the actual emotions. It calculates the accuracy for each emotion separately and then plots these accuracies against the corresponding emotions. Hence it evaluates the accuracy of emotion predictions.

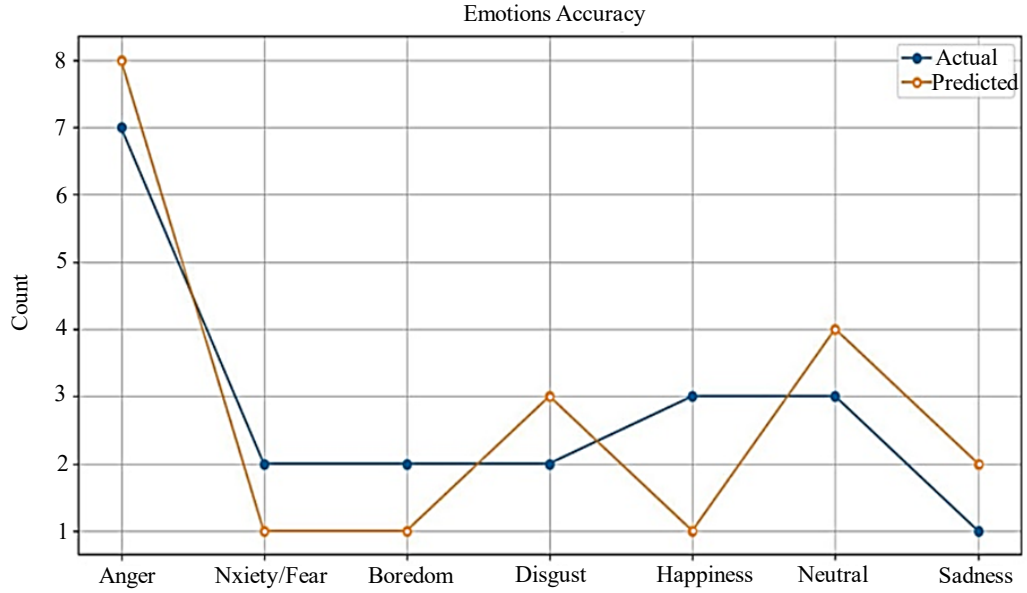

Figure 5. RCNN model.

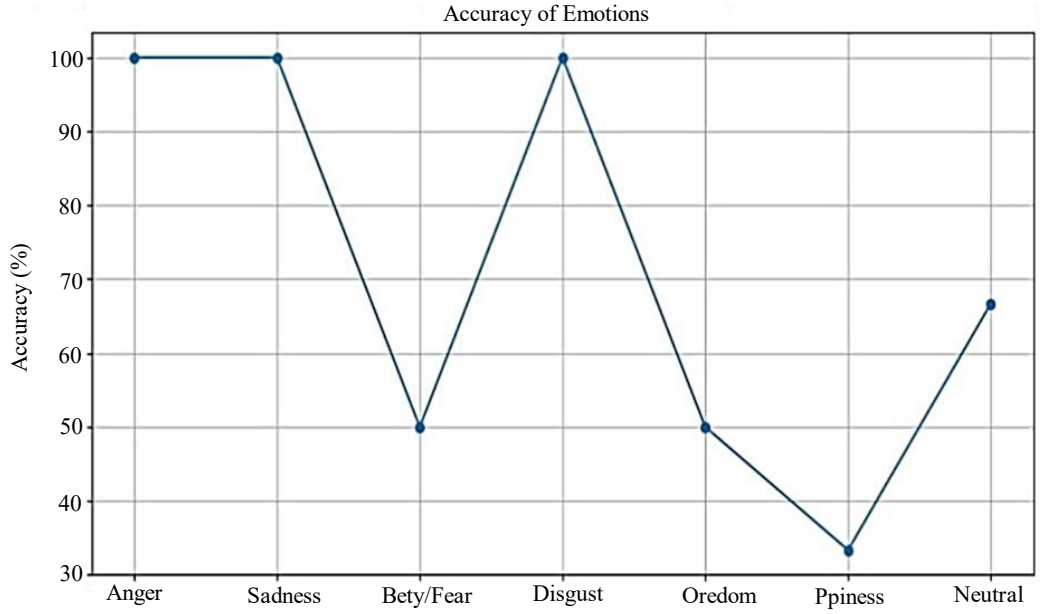

Figure 6. Emotion label accuracy.

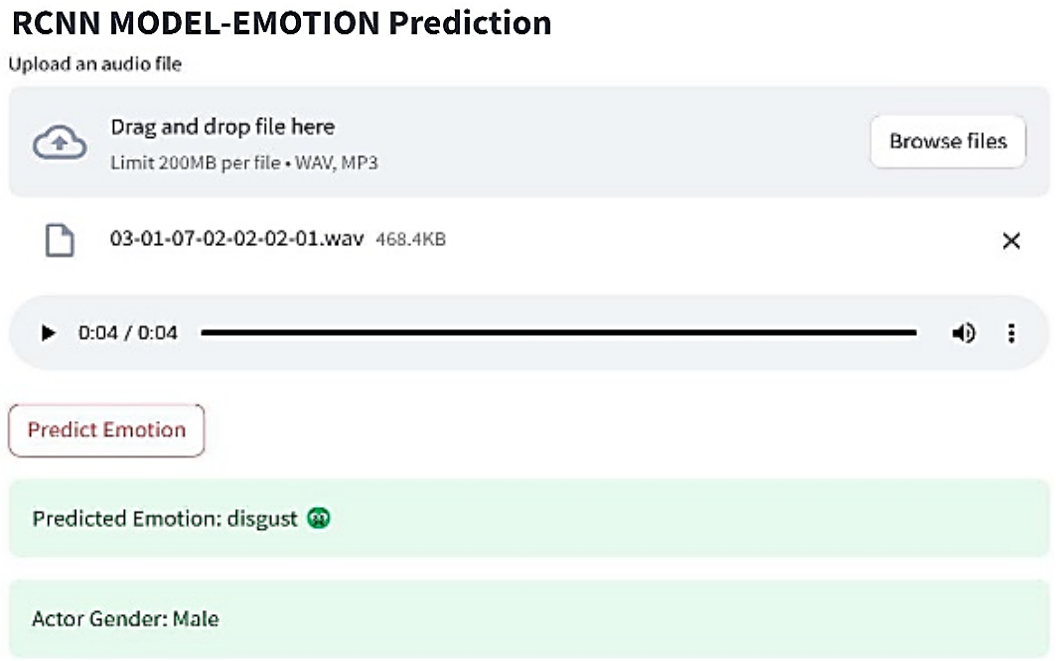

Figure 7. Accuracy of an emotion.

CONCLUSION

In conclusion, using advanced deep learning techniques like Mixed Convolutional Neural Networks (MCNN) and Residual Convolutional Neural Networks (RCNN) can help us understand emotions in speech, especially when we also consider gender information. This approach has the potential to improve how computers interact with humans and how emotional intelligence systems work. By using these neural network architectures and analyzing large and diverse datasets, we can accurately identify emotions in speech while also capturing gender-related signals. This deepens our understanding of emotional expressions and improves the accuracy of emotion recognition. As deep learning technology continues to improve, we anticipate even greater progress in automatically understanding emotions, leading to more reliable systems. Ultimately, using deep learning in speech emotion recognition, particularly when paired with gender data, can transform various fields such as human-computer interaction, emotional understanding, psychology research, and personalized user experiences. These advancements could lead to more empathetic and smoother interactions between humans and machines, making the digital world a friendlier and more emotionally attuned place for everyone. The main discovery is that MCNN models tend to perform better than RCNN models in recognizing emotions from speech.

REFERENCES

1. Singh V, Prasad S. Speech emotion recognition system using gender dependent convolution neural network. Procedia Comput Sci. 2023 Jan 1; 218: 2533–40.
2. Zhang LM, Li Y, Zhang YT, Ng GW, Leau YB, Yan H. A deep learning method using gender-specific features for emotion recognition. Sensors. 2023 Jan 25; 23(3): 1355.
3. Tigga NP, Garg S. Speech Emotion Recognition for multiclass classification using Hybrid CNN-LSTM. International Journal of Microsystems and IoT (IJMIT). 2023; 1: 9–17.
4. Fu H, Zhuang Z, Wang Y, Huang C, Duan W. Cross-Corpus Speech Emotion Recognition Based on Multi-Task Learning and Subdomain Adaptation. Entropy. 2023 Jan 7; 25(1): 124.
5. Wani TM, Gunawan TS, Qadri SA, Kartiwi M, Ambikairajah E. A comprehensive review of speech emotion recognition systems. IEEE Access. 2021 Mar 22; 9: 47795–814.
6. Nicolini M, Ntalampiras S. A Hierarchical Approach for Multilingual Speech Emotion Recognition. In Proceedings of the 12th International Conference on Pattern Recognition Applications and Method (ICPRAM). 2023; 679–685.

7. Latif S, Rana R, Khalifa S, Jurdak R, Epps J, Schuller BW. Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. IEEE Trans Affect Comput. 2020 Apr 1; 13(2): 992–1004.
8. Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden Markov models. Speech Commun. 2003 Nov 1; 41(4): 603–23.
9. Le D, Provost EM. Emotion recognition from spontaneous speech using hidden Markov models with deep belief networks. In 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. 2013 Dec 8; 216–221.
10. Lin YL, Wei G. Speech emotion recognition based on HMM and SVM. In 2005 IEEE international conference on machine learning and cybernetics. 2005 Aug 18; 8: 4898–4901.

