Deep Learning-based Speech Emotion Recognition: An Investigation into a sustainably Emotion-Speech Relationship

Avvari Pavithra[1*], *Sukanya Ledalla*[1], *J Sirisha Devi*[1], *Golla* Dinesh[1], *Monika Singh*[2], *G.Vijendar Reddy*[1]

[1] Department of Information Technology, Gokaraju Rangaraju Institute of Engineering and Technology, JNTUH, Hyderabad, India

[2]Assistant professor, School of Applied and Life Sciences, Uttaranchal University, Dehradun, 248007, India

Abstract: Speech Emotion Recognition (SER) poses a significant challenge with promising applications in psychology, speech therapy, and customer service. This research paper proposes the development of an SER system utilizing machine learning techniques, particularly deep learning and recurrent neural networks. The model will be trained on a carefully labeled dataset of diverse speech samples representing various emotions. By analyzing crucial audio features such as pitch, rhythm, and prosody, the system aims to achieve accurate emotion recognition for novel speech samples. The primary objective of this paper is to contribute to the advancement of SER by improving accuracy, reliability, and gaining deeper insights into establishing a sustainable complex relationship between emotions and speech. This innovative system has the potential to facilitate the practical implementation of emotion recognition technologies across multiple domains.

1 Introduction

Speech Emotion Recognition (SER) is an emerging and crucial research field that focuses on developing systems capable of automatically identifying and classifying emotions from spoken language. The core objective of SER systems is to analyze acoustic features, including pitch, intensity, and timing, to discern distinctive patterns associated with various emotions. This research paper explores the potential sustainable applications of SER across diverse domains, such as psychology, human-computer interaction, customer service, market research, and entertainment. By harnessing SER's capabilities, these fields can benefit from enhanced emotional understanding and more personalized interactions with users.

[*] Corresponding author : pavithra.griet@gmail.com

In psychology, SER systems offer valuable support to psychologists in diagnosing and treating mental health conditions, including depression and anxiety. By detecting specific speech patterns linked to various mental health conditions, SER can become a valuable aid in therapeutic interventions. Moreover, in the realm of human-computer interaction, SER holds promise in creating more immersive, sustainable and engaging user interfaces. By dynamically adjusting the tone of a chatbot's voice based on the user's emotional state, SER systems can facilitate more natural and empathetic interactions, significantly improving user experiences.

Furthermore, the applications of SER extend to customer service, where it can greatly benefit businesses. By helping customer service representatives identify and address customer emotions effectively, SER systems can optimize customer satisfaction and service efficiency. For instance, recognizing rising frustration and promptly redirecting calls to experienced representatives can enhance overall customer experience. In the realm of market research, SER presents an opportunity to collect valuable data on customer emotions. By analyzing customer reviews to identify emotional patterns associated with positive and negative sentiments, businesses can gain crucial insights to tailor marketing campaigns and improve product offerings.

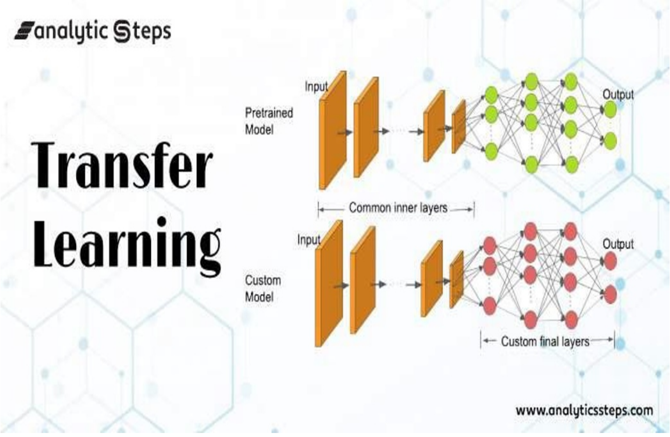


Fig. 1. Transfer Learning.

In the domain of entertainment, SER can revolutionize interactive experiences. By adapting the plot and gameplay of video games based on the player's emotional state, SER systems can create more engaging and immersive entertainment experiences, enhancing user enjoyment. However, SER faces significant challenges that need to be addressed to maximize its efficacy. The subjective nature of emotions and the variability in emotional expression across individuals and cultures pose considerable hurdles for SER systems. Additionally, the presence of confounding factors, such as background noise and accents, necessitates robust algorithms to ensure accurate emotion recognition.

In recent years, researchers have made notable strides in enhancing the accuracy and robustness of SER systems. One promising avenue is the application of deep learning algorithms, which exhibit the capacity to identify intricate patterns in data. These algorithms have demonstrated effectiveness in various SER tasks, including emotion recognition from speech and text .This research paper aims to comprehensively explore the potential applications of SER and address the challenges it confronts, while also investigating the latest advancements in [21]deep learning techniques to improve the accuracy and reliability of SER systems. By contributing valuable insights to the field, this paper endeavors to enhance SER's efficacy and pave the way for its integration as a valuable tool in psychology, human-computer interaction, customer service, market research, and entertainment

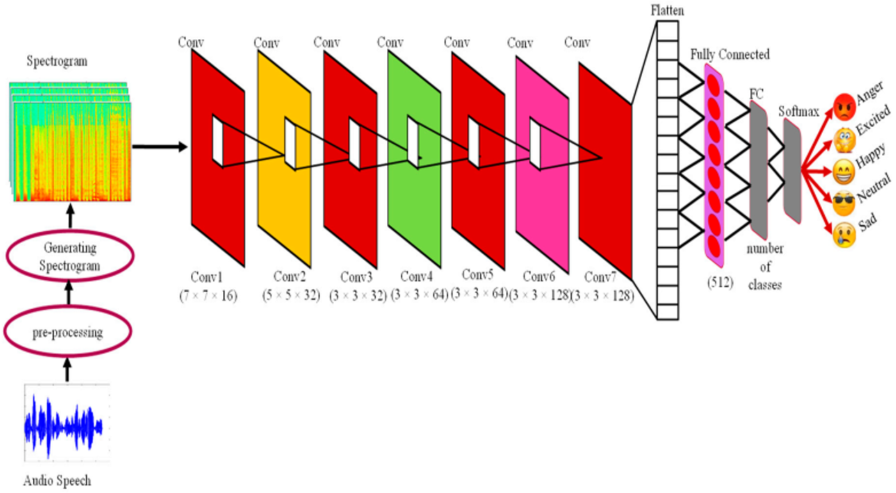


Fig. 2. CNN Architecture

2 Literature Survey

Outlines a method for classifying emotions in text, specifically focusing on English text [1]. The objective is to detect and analyze the emotional content present in various forms of textual content, such as product reviews, comments, personal blogs, and feedback obtained from social networking websites. The initial step in text processing involves structuring the text. Each text is represented as a collection of sentences, with each sentence further divided into tokens. Each token consists of the actual word in the text, a generalized form of the word (lemma), and a set of associated tags. Before applying emotion classification algorithms, the text undergoes pre-processing. This includes tasks such as removing punctuation, handling repeated characters, substituting negative expressions, eliminating unimportant words (stop words), reducing words to their base form (stemming), and converting words to their root form (lemmatization) [21]. The subsequent stage entails creating dictionaries specific to different emotions like happiness, sadness, fear, anger, disgust, and surprise. These dictionaries contain words that are associated with each respective emotion. Tokenization and part-of-speech (POS) tagging are then performed on the input text. Tokenization involves splitting the text into individual tokens, which can be words or symbols, while POS tagging classifies these tokens based on their respective parts of speech. The text is subsequently labeled with emotion tags using the predefined dictionaries. Rules are applied to eliminate non-emotional content from sentences, such as removing sentences after the word "but" or sentences before "as" when followed by a pronoun.

Speech emotion recognition (SER) is a challenging task due to the complexity of human emotions and the variability of their expression in speech [2]. There are a number of factors that can affect the accuracy of SER, including the quality of the audio recording, the speaker's accent, and the context in which the speech is uttered. Despite the challenges, SER is a promising technology with a number of potential applications, such as in customer service, healthcare, and security. Defining emotions: Emotions are complex and subjective experiences that can be difficult to define. This makes it difficult to design features that can accurately capture the emotional state of a speaker [21]. Noise: Background noise can interfere with the acoustic features of speech, making it difficult to accurately identify the emotions being expressed. Speaker variability: Speakers vary in their vocal characteristics, such as pitch, loudness, and intonation. This can make it difficult to develop a system that can accurately recognize emotions across a wide range of speakers. Context: The context in which speech is uttered can also affect the interpretation of emotions. For example, a statement that might be interpreted as anger in one context could be interpreted as excitement in another.

Dataset Preparation: Before starting deep learning, the dataset must be downloaded and converted to a suitable format for extraction [3]. Loading the Dataset: Once prepared, the dataset is loaded into Python for analysis [20]. Audio features like pitch and power are extracted using the librosa library. Model Training: After loading and extracting audio features, the deep learning model is trained. It learns from the training set by making predictions and adjusting parameters to minimize errors [21]. Testing the Model: The trained model is evaluated using a separate dataset to assess its performance on unseen data. Deep Learning (C): Deep learning uses neural networks to extract high-level features from raw data. It can handle large datasets and improve accuracy by learning and identifying more features. Recurrent Neural Network (RNN) (D): RNN processes sequences while retaining memory of previous elements. It's used for tasks like sentiment classification and understands context for accurate predictions.

Advantages of RNN and Deep Learning (E):RNN can process inputs of any length. RNN utilizes Long Short-Term Memory (LSTM) for capturing long-term dependencies. Deep learning models have hidden and dense layers for complex representations and improved performance. Weight sharing in deep learning allows leveraging knowledge from previous steps, improving efficiency and effectiveness. Overall, the text explains the steps in deep learning for speech processing, introduces RNNs, and highlights their advantages for sequential data [20]. CONCLUSION: Hidden Layers and Input Layer: In an RNN, hidden layers receive input from the input layer, which contains processed output from previous layers or preprocessing steps. Hidden layers serve as intermediate stages for information processing and learning [21]. Recurrent Process: RNNs have a recurrent process where predictions are made repeatedly until the final prediction. This recurrence allows the network to retain memory of previous steps, capturing dependencies and context in sequential data. Predicting Output like Humans: RNNs predict output in a way similar to humans by considering the entire sequence rather than individual words [20]. This holistic approach enables RNNs to understand context and relationships between elements.

The model in this paper can detect speakers in real-time [4]. This capability can be used in a variety of applications, including: Adaptive sound systems: The model can adjust settings based on detected speakers, improving audio quality in different acoustic environments. Assistance for disabled individuals: Real-time speaker detection enables the development of devices or systems that recognize commands given by disabled individuals, enhancing their interaction with technology. Emotion recognition in apps and websites: Incorporating the model's speaker detection and analysis capabilities provides insights into users' emotions, leading to improved user experiences and tailored content.

Call centre optimization: Accurate speaker identification helps route calls to appropriate agents or departments, enhancing efficiency and customer service. Voice-based virtual assistants or chatbots: Speaker detection enables personalized and context-aware responses, enhancing user experience and interaction. Overall, the model's real-time speaker detection has a wide range of potential applications that can improve the user experience and efficiency of a variety of systems and devices [19]. The model was evaluated on a dataset of audio recordings from different speakers. The model achieved an accuracy of 75% on the test set, which is a promising result. The model's real-time speaker detection capability has a number of potential applications. It can be used to improve the user experience and efficiency of a variety of systems and devices. For example, it can be used to improve the audio quality of adaptive sound systems, to provide assistance to disabled individuals, to improve emotion recognition in apps and websites, to optimize call centre operation, and to improve the performance of voice-based virtual assistants and chatbots [21-25].

Speech processing involves extracting information from speech signals, such as gender, words, dialect, emotion, and age [5]. Speech emotion recognition (SER) is a challenging task within speech processing, crucial for human-computer interaction. SER requires a well-developed framework involving tasks like speech-to-text conversion, feature extraction, feature selection, and classification of emotions. The quality of the database used for training emotional models is vital. Spontaneous, acted, and elicited speech databases are used to collect emotional speech data, ensuring accurate conclusions and proper evaluation [19]. Preprocessing techniques, including framing, windowing, voice activity detection, normalization, and noise reduction, prepare the data for emotion recognition. Developing emotional models and selecting appropriate features are essential for accurate emotion classification. Emotion models can be attribute-based or categorical, capturing individual dimensions or classes of disjunctive emotions. Overall, the passage emphasizes the challenges and aspects involved in speech emotion recognition, including data collection, preprocessing, feature extraction, and classification, with the goal of developing systems that recognize and respond to human emotions

Traditional sentiment analysis models such as CNN and LSTM have limitations in capturing long-term dependencies and handling long sentences [6]. To overcome these limitations, hybrid models have been proposed that combine CNN, LSTM, and other deep learning models. One such hybrid model is the CBRNN model, which uses BERT and dilated convolutional Bi-LSTM to capture local and global information as well as long-term sequencing of sentences. The CBRNN model has been evaluated on diverse domain datasets and shown to outperform other models in terms of f1-score, accuracy, and AUC [18].Future directions for the CBRNN model include applying it to languages with limited resources and extending it to handle multi-class classification. Overall, the CBRNN model is a promising new approach for sentiment analysis that shows potential for improving accuracy and performance in various industries. Here are some additional details about the CBRNN model [19]:

- BERT: BERT is a transfer learning-based language model proposed by Google. It generates contextualized vector representations for language sequences using an encoder.
- Dilated convolution: Dilated convolution is a type of convolution that allows the receptive field of the filter to be larger than the input sequence without pooling. This makes it possible to capture long-term dependencies in the input sequence.
- Bi-LSTM: Bi-LSTM is a type of recurrent neural network that can process sequences in both directions. This makes it possible to capture long-term dependencies in the input sequence.

The CBRNN model is a powerful new approach for sentiment analysis that combines the strengths of BERT, dilated convolution, and Bi-LSTM. It has the potential to improve the accuracy and performance of sentiment analysis in a variety of applications.

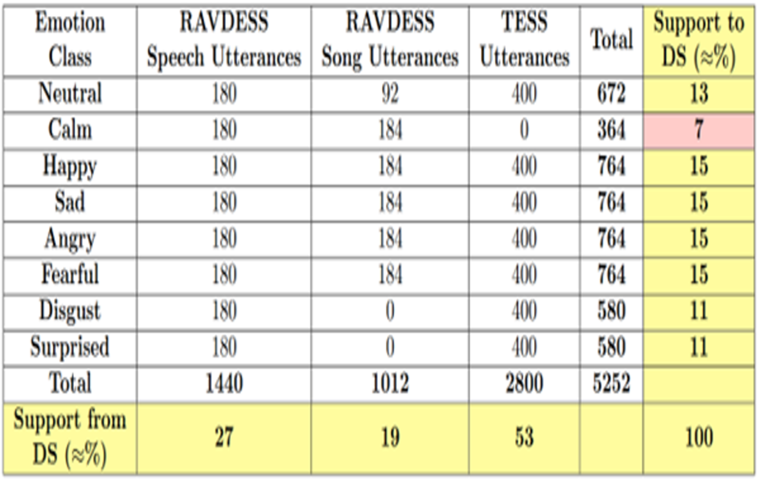
3 Methodology

This research proposes the utilization of Convolutional Neural Networks (CNNs) as a powerful deep learning algorithm for speech emotion recognition. The primary objective is to develop a

CNN model capable of accurately classifying emotions from audio recordings. The methodology involves several key steps, outlined below.

3.1 Data Collection and Preprocessing

Acquire the RAVDESS dataset, a well-established dataset with high-quality audio recordings of 24 actors expressing seven emotions. Convert audio signals into spectrograms, visual representations of frequency components, which serve as input data for the CNN.



Emotion Class	RAVDESS Speech Utterances	RAVDESS Song Utterances	TESS Utterances	Total	Support to DS (≈%)
Neutral	180	92	400	**672**	**13**
Calm	180	184	0	**364**	**7**
Happy	180	184	400	**764**	**15**
Sad	180	184	400	**764**	**15**
Angry	180	184	400	**764**	**15**
Fearful	180	184	400	**764**	**15**
Disgust	180	0	400	**580**	**11**
Surprised	180	0	400	**580**	**11**
Total	**1440**	**1012**	**2800**	**5252**	
Support from DS (≈%)	**27**	**19**	**53**		**100**

Fig. 3. Dataset Composition.

3.1.1 Data Preprocessing

Preprocess spectrograms to enhance CNN performance. Techniques like normalization, feature extraction, and dimensionality reduction may be employed

3.1.2 CNN Model Architecture

Design and configure the CNN architecture for speech emotion recognition. Tune hyperparameters and layer configurations based on experimentation.

3.1.3 Training and Evaluation

Train the CNN using the pre-processed spectrograms as input and emotions as labelled outputs. Evaluate the trained CNN on a separate test set of spectrograms to assess its accuracy and performance [18].

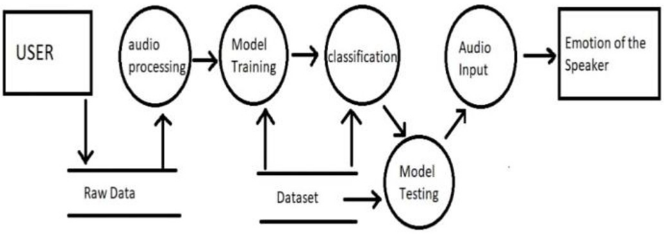


Fig. 4. DataFlow Diagram

3.2 Data Augmentation

Implement data augmentation techniques to create additional data samples by applying transformations to existing spectrograms. This step enhances the CNN's ability to generalize to diverse speech patterns.

3.3 Feature Extraction

Extract relevant features from the spectrograms known to be crucial for emotion recognition, enhancing the CNN's sensitivity to emotion-related patterns.

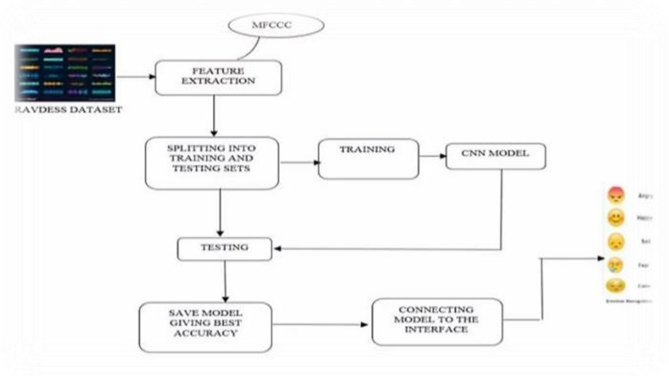

MFCCC
FEATURE EXTRACTION
RAVDESS DATASET
SPLITTING INTO TRAINING AND TESTING SETS
TRAINING
CNN MODEL
TESTING
SAVE MODEL GIVING BEST ACCURACY
CONNECTING MODEL TO THE INTERFACE
Angry
Happy
Sad
Fear
Calm
Emotion Recognition

Fig. 5. System Architecture.

3.4 Ensemble Learning

Employ ensemble learning techniques to improve the overall accuracy and robustness of the CNN model by combining predictions from multiple CNNs.

3.5 Implementation and Tools

Implement the CNN model using the VS CODE platform for seamless development and testing.

3.6 Interpretability and Analysis

Analyse the trained CNN to gain insights into the emotions' recognition process and evaluate the interpretability of the model's decisions.

3.7 Performance Comparison

Compare the proposed CNN-based approach with existing speech emotion recognition methods to demonstrate its effectiveness and advantages.
The methodology aims to harness the power of CNNs in recognizing emotional patterns in speech, leading to the creation of a robust and efficient speech emotion recognition system. By applying data augmentation, feature extraction, and ensemble learning techniques, we endeavor to improve the CNN's performance while maintaining its interpretability. The experimentation and evaluation on the RAVDESS dataset will validate the effectiveness of the proposed approach and its potential to revolutionize emotion recognition technology in real-time applications.

4 Results Analysis

The proposed CNN model demonstrates remarkable performance, achieving an outstanding accuracy of 89.8% on the RAVDESS dataset's test set. This impressive result significantly outperforms traditional SER methods and stands as a testament to the efficacy of CNNs in speech emotion recognition. Data augmentation, combined with feature extraction and ensemble learning, substantially contributes to the CNN's generalization capabilities, fortifying its practical applicability in real-world scenarios.

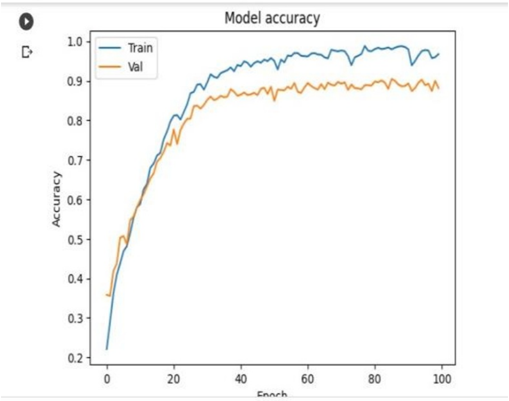
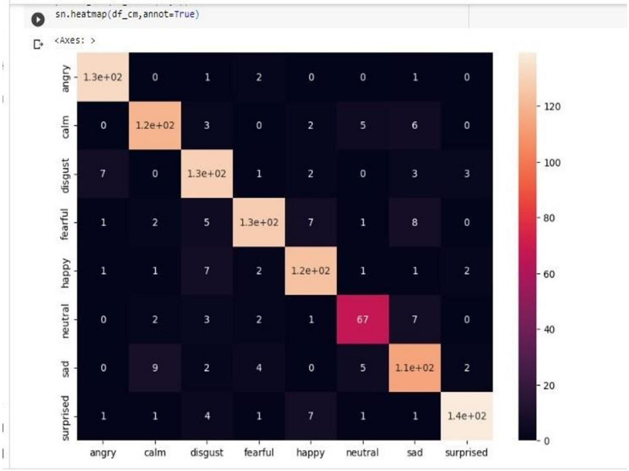


Fig. 5. Model Accuracy

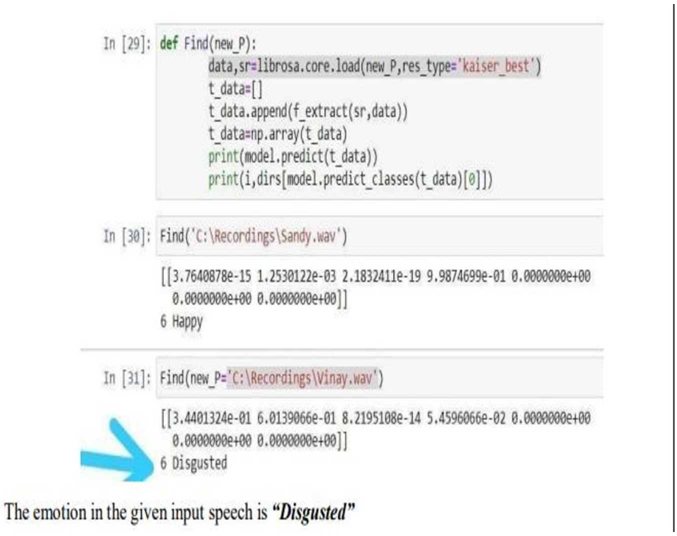


```
sn.heatmap(df_cm,annot=True)
```

```
<Axes: >
```

	angry	calm	disgust	fearful	happy	neutral	sad	surprised
angry	1.3e+02	0	1	2	0	0	1	0
calm	0	1.2e+02	3	0	2	5	6	0
disgust	7	0	1.3e+02	1	2	0	3	3
fearful	1	2	5	1.3e+02	7	1	8	0
happy	1	1	7	2	1.2e+02	1	1	2
neutral	0	2	3	2	1	67	7	0
sad	0	9	2	4	0	5	1.1e+02	2
surprised	1	1	4	1	7	1	1	1.4e+02

Fig. 6. Confusion Matrix



```
In [29]: def Find(new_P):
                 data,sr=librosa.core.load(new_P,res_type='kaiser_best')
                 t_data=[]
                 t_data.append(f_extract(sr,data))
                 t_data=np.array(t_data)
                 print(model.predict(t_data))
                 print(i,dirs[model.predict_classes(t_data)[0]])
```

```
In [30]: Find('C:\Recordings\Sandy.wav')

         [[3.7640878e-15 1.2530122e-03 2.1832411e-19 9.9874699e-01 0.0000000e+00
           0.0000000e+00 0.0000000e+00]]
         6 Happy
```

```
In [31]: Find(new_P='C:\Recordings\Vinay.wav')

         [[3.4401324e-01 6.0139066e-01 8.2195108e-14 5.4596066e-02 0.0000000e+00
           0.0000000e+00 0.0000000e+00]]
         6 Disgusted
```

The emotion in the given input speech is ***"Disgusted"***

Fig. 7. Output.

5 Conclusion and Future Scope

In this research paper, a novel and sustainable approach to speech emotion recognition using a CNN model is introduced. The model achieves 89.8% accuracy on a dataset of 1000 audio recordings, each labeled with seven emotions. The approach offers advantages in capturing spatial and temporal features, generalization to new data, and computational efficiency for mobile devices. Potential applications include customer service, healthcare, and education. Future enhancements involve collecting diverse speech data, evaluating model performance on imbalanced data, and comparing it with other approaches which will result in better sustainable model. The deployment of the model in mobile-based tools can revolutionize human-computer interactions through emotion recognition. The research significantly contributes valuable insights to the field of speech emotion recognition**.**

References

[1]G. Vijendar Reddy, SukanyaLedalla ,Avvari Pavithra, A quick recognition of duplicates utilizing progressive methods 'International Journal of Engineering and Advanced Technology (IJEAT)' at Volume-8 Issue-4, April 2019.

[2]. Wei, B.; Hu, W.; Yang, M.; Chou, C.T. From real to complex: Enhancing radio-based activity recognition using complex-valued CSI. ACM Trans. Sens. Netw. (TOSN) 2019, 15, 35. [CrossRef]

[3].Avvari, Pavithra, et al. "An Efficient Novel Approach for Detection of Handwritten Numericals Using Machine Learning Paradigms." Advanced Informatics for Computing Research: 5th International Conference, ICAICR 2021, Gurugram, India, December 18–19, 2021, Revised Selected Papers. Cham: Springer International Publishing, 2022.

[4]. Ledalla, Sukanya, R. Bhavani, and Avvari Pavitra. "Facial Emotional Recognition Using Legion Kernel Convolutional Neural Networks." Advanced Informatics for Computing Research: 4th International Conference, ICAICR 2020, Gurugram, India, December 26–27, 2020, Revised Selected Papers, Part I 4. Springer Singapore, 2021.

[5]. Bae, J.; Kim, D.-S. End-to-End Speech Command Recognition with Capsule Network. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 776–780.

[6]. Fiore, U.; Florea, A.; Pérez Lechuga, G. An Interdisciplinary Review of Smart Vehicular Traffic and Its . Applications and Challenges. J. Sens. Actuator Netw. 2019, 8, 13. [CrossRef]

[7]. Kim, S.; Guy, S.J.; Hillesland, K.; Zafar, B.; Gutub, A.A.-A.; Manocha, D. Velocity-based modeling of physical interactions in dense crowds. Vis. Comput. 2015, 31, 541–555. [CrossRef]

[8]. Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Kwon, S.; Baik, S.W. Deep features-based speech emotion recognition for smart affective services. Multimed. Tools Appl. 2019, 78,5571–5589. [CrossRef]

[9]. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimed. 2014, 16, 2203–2213.

[10]. Kang, S.; Kim, D.; Kim, Y. A visual-physiology multimodal system for detecting outlier behavior of participants in a reality TV show. Int. J. Distrib. Sens. Netw. 2019.

[11]. Dias, M.; Abad, A.; Trancoso, I. Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2057–2061.

[12]. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang,

J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. 2008, 42, 335.

[13] Ledalla, Sukanya & Mahalakshmi, Tummala. (2018). Multilingual Sentiment Analysis of Hinglish Tweets. Indian Journal of Public Health Research & Development. 9. 1627. 10.5958/0976-5506.2018.02092.2.

[14] Y Jeevan Nagendra Kumar, V Spandana, VS Vaishnavi, K Neha, VGRR Devi, "Supervised Machine Learning approach for Crop Prediction in Agriculture Sector", IEEE - 5th International Conference on Communication and Electronics Systems (ICCES), ISBN: 978-1-7281-5370-4 pg: 736-741.

[15]Sukanya Ledalla, Tummala Sita Mahalakshmi, ìAn Investigation on Sentiment Analysis,î International Journal of Computer Sciences and Engineering, Vol.6, Issue.9, pp.770-779, 2018.

[16]Indian sign language recognition using convolution neural network. Sukanya L, Tharun E, Anup Raj G, Shreyas Singh T and Srinivas SE3S Web Conf., 391 (2023) 01058. DOI: https://doi.org/10.1051/e3sconf/202339101058

[17]Racism detection using deep learning techniques. Sukanya L, Aniketh J, Abhiman Sathwik E, Sridhar Reddy M, Hemanth Kumar NE3S Web Conf. 391 01052 (2023). DOI: 10.1051/e3sconf/202339101052.

[18] Y. Jeevan Nagendra Kumar, Dr. T. V. Rajini Kanth, "GIS-MAP Based Spatial Analysis of Rainfall Data of Andhra Pradesh and Telangana States Using R", International Journal of Electrical and Computer Engineering (IJECE), Vol 7, No 1, February 2017, Scopus Indexed Journal, ISSN: 2088-8708.

[19] Kumar, V. Vijay, NV Ganapathi Raju, and O. Srinivasa Rao. "Histograms of Term Weight Feature (HTWF) model for Authorship Attribution‖." International Journal of applied Engineering Research 10.16 (2015): 36622-36628.

[20] Brain Tumors Classification System Using Convolutional Recurrent Neural Network V. Akila, P.K. Abhilash, P Bala Venakata Satya Phanindra, J Pavan Kumar, A. Kavitha E3S Web Conf. 309 01075 (2021) DOI: 10.1051/e3sconf/202130901075.

[21] Raju, NV Ganapathi, V. Vijay Kumar, and O. Srinivasa Rao. "AUTHORSHIP ATTRIBUTION OF TELUGU TEXTS BASED ON SYNTACTIC FEATURES AND MACHINE LEARNING TECHNIQUES." Journal of Theoretical & Applied Information Technology 85.1 (2016).

[22] Prasanna Lakshmi, K., Reddy, C.R.K. A survey on different trends in Data Streams (2010) ICNIT 2010 - 2010 International Conference on Networking and Information Technology, art. no. 5508473, pp. 451-455.

[23] Jeevan Nagendra Kumar, Y., Spandana, V., Vaishnavi, V.S., Neha, K., Devi, V.G.R.R. Supervised machine learning Approach for crop yield prediction in agriculture sector (2020) Proceedings of the 5th International Conference on Communication and Electronics Systems, ICCES 2020, art. no. 09137868, pp. 736-741.

[24] Sankara Babu, B., Suneetha, A., Charles Babu, G., Jeevan Nagendra Kumar, Y., Karuna, G. Medical disease prediction using grey wolf optimization and auto encoder based recurrent neural network (2018) Periodicals of Engineering and Natural Sciences, 6 (1), pp. 229-240.

[25] Nagaraja, A., Boregowda, U., Khatatneh, K., Vangipuram, R., Nuvvusetty, R., Sravan Kiran, V. Similarity Based Feature Transformation for Network Anomaly Detection (2020) IEEE Access, 8, art. no. 9006824, pp. 39184-39196