MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

Project Report

on

Speech Emotion Recognition using CNN-TRANSFORMER Architecture

Submitted By:

Sarthak Mangalmurti

0901AM211051

Ojshav Saxena

0901AM211035

Faculty Mentor:

Dr. Tej Singh

Assistant Professor

Centre for Artificial Intelligence

CENTRE FOR ARTIFICIAL INTELLIGENCE

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

JULY-DEC. 2023

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

CERTIFICATE

This is certified that **Sarthak Mangalmurti**(0901AM211051), **Ojshav Saxena**(0901AM211035) has submitted the project report titled **Speech Emotion Recognition using CNN-TRANSFORMER Architecture** under the mentorship of **Dr. Tej Singh**, in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in **Artificial intelligence and Machine Learning** from Madhav Institute of Technology and Science, Gwalior.

23/11/23

Dr. Tej Singh
Faculty Mentor
Assistant Professor
Centre for Artificial Intelligence

23/11/23

Dr. R. R. Singh
Coordinator
Centre for Artificial Intelligence

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Artificial intelligence and Machine Learning at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of Dr. Tej Singh, Assistant Professor, Centre for Artificial Intelligence

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Sarthak Sarthak Mangalmurti
0901AM211051
3rd Year,
Centre for Artificial Intelligence

Saxena Ojshav Saxena
0901AM211035
3rd Year,
Centre for Artificial Intelligence

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

NAAC Accredited with A++ Grade

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Centre for Artificial Intelligence,** for allowing me to explore this project. I humbly thank **Dr. R. R. Singh**, Coordinator, Centre for Artificial Intelligence, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Dr. Tej Singh**, Assistant Professor, Centre for Artificial Intelligence, for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Sarthak Mangalmurti
0901AM211051
3rd Year,
Centre for Artificial Intelligence

Ojshav Saxena
0901AM211035
3rd Year,
Centre for Artificial Intelligence

TABLE OF CONTENTS

ABSTRACT

In the past few years, recognition of emotions from speech has become increasingly popular, due to its wide-ranging potential applications across various fields such as healthcare, and entertainment. The capability to identify and understand human emotions through speech is an intriguing and important research area, offering significant implications for a variety of practical uses. Sentiment analysis has garnered considerable attention due to its potential to enhance human-computer interaction, create more empathetic AI systems, and improve the quality of healthcare and entertainment experiences. In the contemporary era, Speech Emotion Recognition (SER) is increasingly vital, with its applications spanning across human-computer interactions for more empathetic AI, early detection of mental health issues, personalized content and education, market research insights, content creation, and improved accessibility. SER is paramount in addressing the evolving needs of technology, healthcare, education, and entertainment, contributing to a more inclusive, emotionally intelligent, and interconnected digital world.

Keywords— Speech Recognition, CNN, Transformer

सारः

पिछले कुछ वर्षों में, भाषा से भावनाओं की पहचान का प्रमाण बढ़ रहा है, जिसका कारण इसे स्वास्थ्य, और मनोरंजन जैसे विभिन्न क्षेत्रों में व्यापक अनुप्रयोग की संभावना है। मानव भाषा के माध्यम से भावनाओं की पहचान और समझ की क्षमता, एक रोचक और महत्वपूर्ण अनुसंधान क्षेत्र है, जिसमें विभिन्न व्यावसायिक उपयोगों के लिए महत्वपूर्ण परिणाम हैं। भावना विश्लेषण ने बड़ी ध्यानाकर्षण प्राप्त किया है क्योंकि इसमें मानव-कंप्यूटर इंटरएक्शन को बढ़ावा देने, अधिक सहानुभूति वाले ए.आई. सिस्टम्स बनाने और स्वास्थ्य और मनोरंजन अनुभवों की गुणवत्ता में सुधार करने की क्षमता है। समकालीन युग में, भाषा भावना पहचान (एस ई आर) बढ़ती हुई है, जिसके अनुप्रयोग मानव-कंप्यूटर इंटरएक्शन के लिए और अधिक सहानुभूति वाले ए.आई. के लिए, मानसिक स्वास्थ्य समस्याओं की पहली पहचान के लिए, व्यक्तिगत सामग्री और शिक्षा के लिए, बाजार अनुसंधान अंशों के लिए, सामग्री निर्माण के लिए, और सुधारित पहुँच के लिए फैले हुए हैं। एस ई आर, प्रौद्योगिकी, स्वास्थ्य, शिक्षा, और मनोरंजन की बदलती आवश्यकताओं का समाधान करने में महत्वपूर्ण है, जो एक समृद्धि, भावनात्मक बुद्धिमत्ता, और एक संबद्ध डिजिटल दुनिया की दिशा में योगदान कर रहा है।

कीवर्ड— भाषा पहचान, सीएनएन, ट्रांसफॉर्मर

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: PROJECT OVERVIEW

1.1. Introduction

Speech Emotion Recognition (SER) has become increasingly popular for several reasons. Emotions are a fundamental aspect of human communication, and understanding them is crucial in various domains. When humans communicate, they convey not only the content of their messages but also their emotional state. SER provides a means to capture this emotional content, which is often subtle and nuanced, enriching the interactions between humans and machines. It has the potential to enhance the user experience in applications like virtual assistants, customer service, and entertainment by enabling more empathetic and responsive interactions. SER using audio data is particularly significant for several reasons. Unlike text-based SER, audio data contains not only the words spoken but also crucial prosodic features like tone, pitch, and timing, which play a vital role in conveying emotions. Additionally, audio-based SER is essential in case we do not have text data, or have very little of it such as analysing emotional states in non-textual content like podcasts, phone calls, and voice messages. As a result, audio-based SER complements text-based SER and expands the scope of emotion recognition.

However, SER comes with its fair share of challenges. Emotions are complex and multifaceted, and their expression in speech varies greatly among individuals and across different languages and cultures. Noise, accent, and background sounds can further complicate the recognition process. Additionally, the emotional states of individuals can be highly contextual, making it challenging to achieve consistent recognition accuracy across diverse scenarios. Overcoming these challenges requires the development of robust models and the utilization of large, diverse datasets.

1.2. Objectives and Scope

The recognition of emotions in speech is essential for monitoring psychological health and well-being, which has gained significant attention in recent years. SER can be used to detect early signs of cognitive problems by analysing changes in emotional states through voice data. It enables remote patient monitoring and timely interventions, especially in telehealth scenarios, contributing to improved mental health outcomes. In the era of data-driven personalization, SER plays a vital role in tailoring content, services, and recommendations to individual preferences and emotional states. For example, in the entertainment industry, it can help streaming platforms suggest movies, music, or content that aligns with the viewer's mood, creating a more engaging and customized user experience. With the growth of online education and e-learning platforms, SER can enhance the effectiveness of remote teaching by gauging student engagement and emotional responses during virtual classes. Educators can use this data to adjust their teaching methods, identify struggling students, and offer additional support, ultimately improving the quality of online education. SER is a valuable tool in market research and customer sentiment analysis, particularly in the era of social media and online reviews. It enables companies to make data-driven decisions to improve their offerings and address customer concerns. In the digital content creation landscape, SER can be utilized to assess the emotional impact of multimedia content, such as video ads or social media campaigns. This data can guide content creators in producing more emotionally resonant and engaging material, contributing to the success of marketing campaigns and online content. SER is essential for making technology more accessible and inclusive. It can benefit individuals with speech and language disabilities by allowing them to convey their emotions and intentions through speech, thereby providing a means for effective communication and interaction with technology and the digital world. In an increasingly globalized world, SER is vital for recognizing emotions across different languages and cultural contexts. It can be used to adapt content and services to various linguistic and cultural nuances, ensuring that emotional recognition remains relevant and effective in diverse settings.

1.3. Project Features

This paper introduces an Emotion Recognition System that harness deep learning to precisely detect emotions from speech signals. Our methodology focuses on transforming raw audio data into informative Mel spectrogram representations[1]. We propose the "CNN-TM" architecture, a fusion of parallel Convolutional Neural Networks (CNNs) [2] and a Transformer encoder, designed to capture emotional cues from speech spectrograms [1]. This synergy is poised to revolutionize SER, allowing for a more nuanced and context-aware understanding of emotional content within speech.

1.4. Feasibility

The feasibility of our project is supported by the supervised learning approach, where our ERS is trained to minimize cross-entropy loss, optimizing over 150 epochs with an Adam optimizer[3]. To prevent overfitting, we incorporate dropout layers and utilize the Rectified Linear Unit (ReLU) activation function for introducing non-linearity[4]. The model undergoes rigorous evaluation using precision, recall, F1-Score, and accuracy metrics, supported by the confusion matrix, to assess its classification capabilities.

1.5. System Requirements

All experiments and training procedures have been conducted on a laptop equipped with a 2.5 GHz Dual-Core Intel® Core i5 processor, 8 GB of memory, and a 512 GB SSD hard drive. Despite the relatively modest hardware specifications, the model demonstrates its capability to effectively recognize emotions in speech signals. And Kaggle notebook with GPU100 Accelerator.

CHAPTER 2: LITERATURE REVIEW

Speech Emotion Recognition (SER) has evolved as a dynamic and evolving field at the intersection of speech processing, machine learning, and human-computer interaction. This section provides an in-depth exploration of key research and developments in SER, shedding light on its foundational concepts and the latest advancements. There have been various different approaches for sentiment analysis used by different researchers over the period of time. In [6], the research introduces the concept of an Audio Spectrogram Transformer (AST), which is a model designed for Speech Emotion Recognition (SER) and AST is based on the transformer architecture [7] relies solely on attention mechanism to analyse audio spectrograms [1]. The paper proposes a method for transferring knowledge from a Vision Transformer (ViT) pretrained on ImageNet, which proves to be highly beneficial in enhancing AST's performance, making it a promising development in the field of SER.

The researchers proposed a framework where audio representations are learned with guidance from in the realm of audiovisual speech, researchers utilize a generative training technique focused on the visual modality in [8]. This method involves animating a static image to synchronize with an audio clip, and then fine-tuning the resulting video to closely resemble the authentic video. The researchers introduced PyNADA, featuring a novel activation function called ADA in [9]. Notably, a single PyNADA neuron can successfully learn the XOR logic due to the unique apical dendrite activation. The research demonstrates that ADA and PyNADA outperform standard neurons using ReLU [4] and leaky ReLU [10] activations across various tasks and neural architectures. This research takes inspiration from the recent study done in [11], showcasing the power of neural networks [12].

The research introduces SPEL that enhances the performance of machine learning models when combined into an ensemble in [13]. Instead of merely aggregating models, this approach allows them to iteratively learn from each other, improving individual models and transferring knowledge about the target domain. The study demonstrates the effectiveness of SPEL through experiments on three audio tasks, outperforming baseline ensemble models. Importantly, it illustrates that applying self-paced learning individually to models is less effective, emphasizing the value of models learning collaboratively within an ensemble. Additionally, ablation results on the CREMA-D dataset reaffirm the benefits of this knowledge-sharing approach[5].

CHAPTER 3: PRELIMINARY DESIGN

In this work we propose an emotion recognition system using the spectrogram of speech signals of different emotions, using two parallel convolutional neural network (CNNs) and a transformer encoder, Each CNN has three convolutional layers and two pooling layers. The outputs of the two CNNs are concatenated and fed into the transformer encoder. The transformer encoder has six layers, each with a multi-head attention mechanism. The output of the transformer encoder is fed into a fully connected layer to predict the emotion. In this approach the model is trained using a supervised learning approach. The training dataset consists of pairs of speech spectrograms and emotion labels. The model is the training process involves minimizing the cross-entropy loss between the predicted emotion labels and the actual ground truth labels. To infer the emotion of a new speech signal, the input speech signal is converted to a spectrogram and fed into the model. The model predicts the most likely emotion based on the output of the fully connected layer.

3.1. Model Architecture

Our Emotion Recognition System (ERS) employs a unique model architecture known as the "CNN-TM" designed to effectively capture emotional cues from speech signals. This architecture integrates two fundamental components: parallel Convolutional Neural Networks (CNNs) and a Transformer encoder. Let's delve deeper into each component:

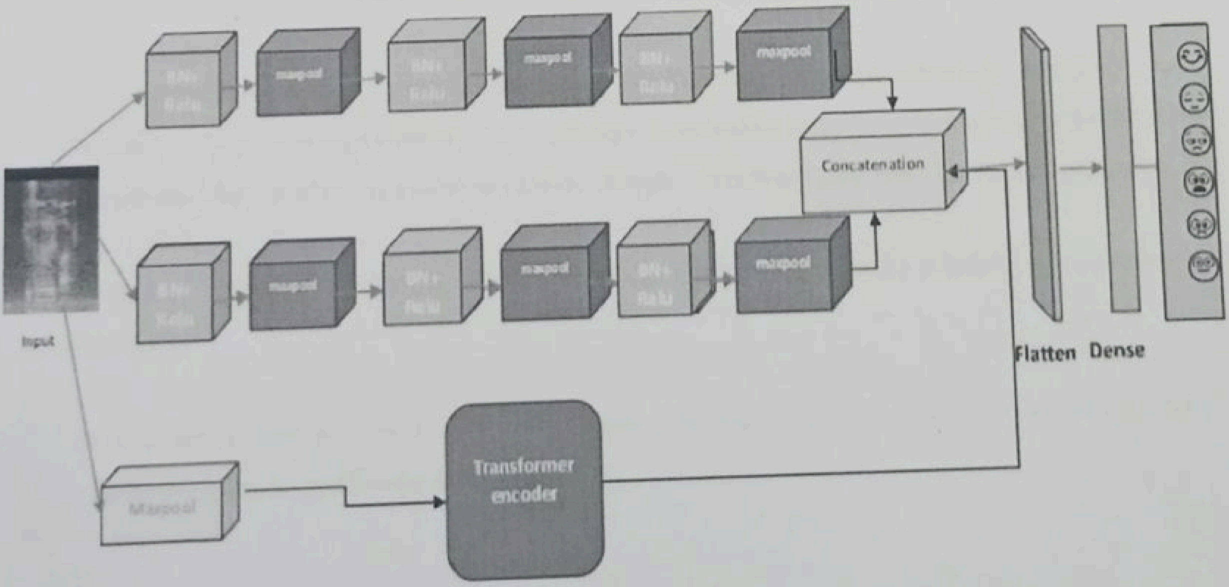


Figure 3. Model Architecture

3.1.1. Parallel CNN Blocks:

The CNN-TM features two parallel CNN blocks, each designed to extract distinct features from the input speech spectrogram. These CNN blocks enable understanding complex patterns and emotional cues within the data. Each CNN block comprises three key elements: convolutional layers, batch normalization, ReLU activation functions, and max-pooling layers. The initial convolutional layer acts as a feature extractor, identifying low-level features within the spectrogram data. The output from this layer is passed through batch normalization to normalize the feature maps and enhance training stability. ReLU activation functions introduce non-linearity to the model.

Max-pooling layers are strategically placed to reduce the spatial dimensions of the feature maps. This reduces computational complexity while preserving essential features. The three convolutional layers within each block work in tandem to capture features of varying complexity. The output of these layers is a collection of high-level features that are distinct and informative. By employing two parallel CNN blocks, the model can focus on different aspects of the spectrogram simultaneously, enhancing its overall feature extraction capabilities.

3.1.2. Transformer Encoder:

The Transformer encoder component plays a pivotal role in capturing complex patterns and relationships within the spectrogram data. This encoder consists of six layers, each incorporating multi-head attention mechanisms and feedforward networks.

Within each layer of the Transformer encoder, there exists a multi-head attention system that enables self-attention. This mechanism empowers the model to assess the significance of different segments within the spectrogram when making predictions. Simultaneously considering various aspects, the model can acquire a deeper understanding of complex relationships present in the data.

Feedforward Networks: After self-attention, feedforward networks introduce additional non-linearity. These networks consist of multiple layers, including linear transformations and activation functions. They enable the model to capture complex relationships within the spectrogram data.

Output of Transformer Encoder: The output of the Transformer encoder is a rich representation of the input

data, enriched with emotional cues and patterns. This representation is then processed by a fully connected layer to predict the emotion.

3.1.3. Fully Connected Layer:

The final layer of the model architecture is responsible for processing the output of the transformer encoder and predicting the emotion category. The fully connected layer is equipped with a SoftMax [14] activation function, which converts the model's output into probability scores for each emotion category. The category with the highest probability score is recognised to be the predicted emotion, allowing the model to effectively classify speech signals into one of the predefined emotional states.

In summary, the CNN-TM architecture is designed to effectively capture emotional cues from speech spectrograms. It combines parallel CNN blocks for feature extraction with a Transformer encoder for capturing complex patterns and relationships. The model's fully connected layer handles emotion prediction, making it a powerful tool for Speech Emotion Recognition (SER). The model's architecture has been fine-tuned and optimized to enhance its accuracy and generalization capabilities, contributing to its robust performance in recognizing emotions in speech signals.

3.2. Model Training

The training of our ERS is a critical phase in achieving accurate emotion recognition. The following aspects characterize our model training process:

3.2.1 Supervised Learning

Our approach adopts a supervised learning paradigm. The training dataset is composed of pairs of speech spectrograms and emotion labels. The model's objective is to to minimize the cross-entropy loss, the goal is reducing discrepancies among the predicted emotions and the actual ground truth labels.

3.2.2 Training Parameters

- Epochs: Our model undergoes 150 training epochs. Each epoch represents a full iteration through the training data, allowing the model to gradually improve its performance.
- Loss Function: We employ the Cross Entropy loss function. This function quantifies the discrepancy between predicted emotion distributions and actual emotion labels.
- Optimizer: Our choice of optimizer is AdamW, which combines the benefits of adaptive optimization and weight decay. It helps fine-tune model parameters.
- Learning Rate: we have taken 1e-4 as our learning rate, ensuring a balance between fast convergence and stability during training.

3.2.3 Training Loop

The training loop involves an iterative process in which the model passes through the training dataset, computes loss, and updates its weights to minimize the loss. This process is repeated over multiple epochs to progressively improve the model.

3.3 Dropout and Activation Function

During training, we employ dropout layers with a probability of 0.4. Dropout is essential for preventing overfitting, as it randomly sets a portion of neurons' outputs to During the training process, the utilization of stochastic regularization aims to drive certain model activations to zero. This technique encourages the model to enhance its learning capabilities. robust and generalizable features. The activation function used throughout the model architecture is Rectified Linear Unit (ReLU).

3.4 SoftMax Function

The SER model's SoftMax function is an essential part. It converts the unprocessed output of the model into probability distributions across various emotional states. To put it simply, the softmax function gives each emotion class a probability score so that the model can decide which emotions to classify. In order to translate the model's learnt characteristics into a format that can be understood for emotion recognition, the SoftMax function is essential.

The SoftMax function converts an input vector of real numbers, represented as Z, where Z = [z1, z2,..., zk], into a probability distribution, represented as P = [p1, p2,..., pk]. The probability that an input belongs to class i is represented by each Pi. This is how the SoftMax function is expressed mathematically:

$$\mathrm{P_i} = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}}$$

Equation 1

In Eq (1), Pi represents the likelihood that the input is a member of class i. zi is the input vector's ith element.

CHAPTER 4: FINAL ANALYSIS AND DESIGN

4.1 Dataset Selection and Description

Our research capitalizes on the CREMA-D dataset, a pivotal resource in the field of Speech Emotion Recognition (SER). This dataset offers a diverse collection of audio recordings, each annotated with primary emotion labels. These primary labels correspond to six distinct emotions, namely sadness, anger, disgust, fear, happiness, and neutrality. The controlled conditions under which actors express emotions make CREMA-D a reliable resource for emotional speech analysis.

4.2 Feature Extraction

Feature extraction plays a pivotal role in SER, transforming raw audio data into informative representations that capture relevant emotional content. In our research, we employed the following feature extraction pipeline:

4.2.1 Mel Spectrogram Representation:

We extracted Mel spectrogram features using the 'torchaudio' library. The Mel spectrogram [1] is a well-established feature representation for speech analysis and SER. It effectively captures the spectral content of audio signals, aligning with human auditory perception. In Fig. 2, we can see spectrogram images for different emotion which are used as an input for our model

- Raw Audio Data: We began by loading audio samples from the CREMA-D dataset, resulting in raw audio waveforms and their respective sample rates.
- Conversion to PyTorch Tensors: For efficient processing, we converted audio waveforms into PyTorch tensors. This step allowed us to harness available hardware resources, whether CUDA GPU or CPU, for computation.
- Mel Spectrogram Computation: The core of our feature extraction involved computing Mel spectrograms. This process involved fine-tuning parameters such as the number of Mel filterbanks, analysis window size, and hop size between windows. These settings influenced the level of detail captured in our spectrogram representations.
- Power Spectrogram to Decibel (dB) Scale: To enhance the informativeness of our spectrogram, we applied the conversion to a decibel (dB) scale. This transformation facilitated the capture of subtle changes in audio intensity, a crucial factor in recognizing emotional expression in speech.
- Data Organization: Throughout this feature extraction process, we accumulated these Mel spectrogram representations into a list, serving as a central repository for these transformed features.

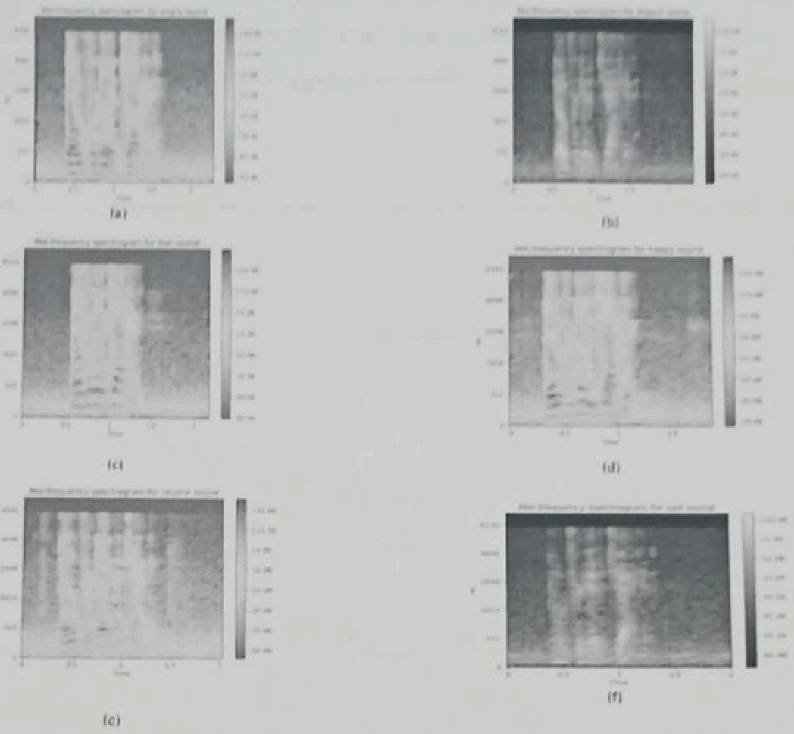


Figure2.Spectogram Images of all Emotions

4.3 Data Splitting and Preprocessing

The first step in our experimental process was to divide the dataset into three distinct categories: training, validation, and testing. We employed a stratified approach to maintain class balance, ensuring that each emotional state was well-represented in each split. Our final split resulted in a training set of 4,688 samples, a validation set of 521 samples, and a test set of 2,233 samples. This careful division of data provided a solid foundation for model training and evaluation.

4.4 Model Training and Evaluation

We adopted the PyTorch framework to develop and train our Speech Emotion Recognition (SER) model. The model was trained over 150 epochs, with the primary objective of minimizing the classification loss. As seen in Fig. 3, The model's training progress was monitored by examining the training loss and accuracy metrics. The "training loss" indicated the model's ability to reduce discrepancies between its predictions and true labels, while the "training accuracy" measured its correctness in classifying examples within the training set.

The validation dataset served as an independent benchmark to assess the model's generalization capabilities. The "validation loss" measure the model's precision in predicting emotional labels on unseen data, while the "validation accuracy" gauged its ability to generalize to new, previously unseen samples.

We introduced an "early stopping mechanism" to mitigate overfitting, ensuring that the model maintained its performance on new data.

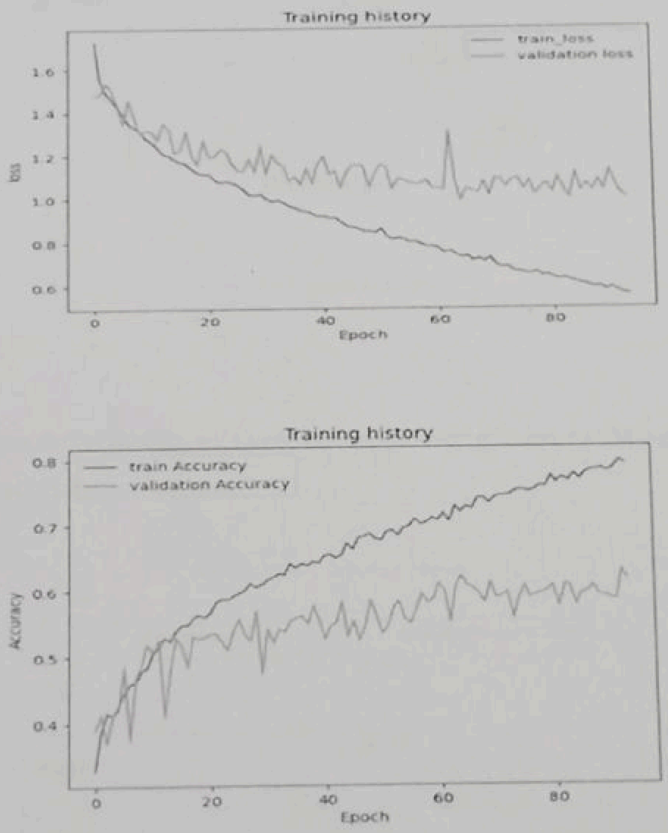
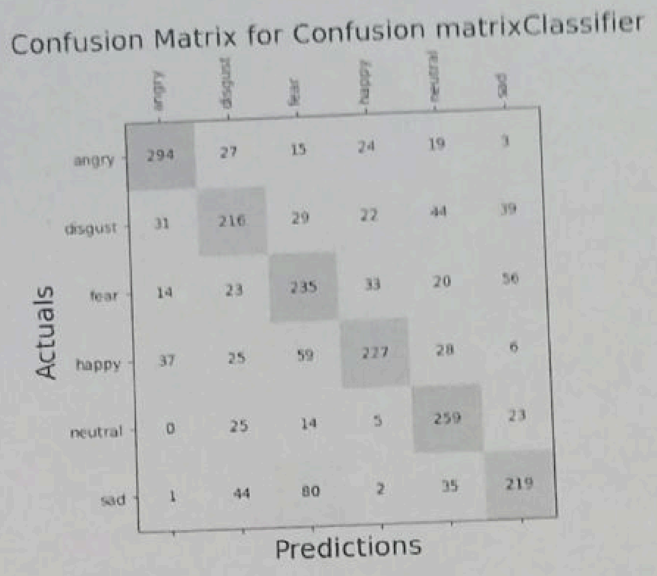


Figure 5.Training History for Accuracy and Loss

4.5 Confusion Matrix and Most Confusing Classes

To delve further into the model's classification outcomes, As seen in Fig. 4, we employed a "confusion matrix" to visualize the results. This matrix presented not only the instances that were correctly classified but also the instances where misclassifications occurred for each emotional category. It offered a comprehensive view of the model's classification accuracy and performance.



	angry	disgust	fear	happy	neutral	sad
angry	294	27	15	24	19	3
disgust	31	216	29	22	44	39
fear	14	23	235	33	20	56
happy	37	25	59	227	28	6
neutral	0	25	14	5	259	23
sad	1	44	80	2	35	219

Figure 6.Confusion Matrix

When analyzing the confusion matrix, we identified the following "most confusing classes" along with their respective misclassification rates:

Table I. Most Confusing Classes

Emotions	*Misclassification rates*
Sad	13.56%
Happy	23.31%
Disgust	15.17%
Fear	23.46%
Angry	12.39%
neutral	25.41%

These insights provided a deeper understanding of the emotional states where the model faced challenges, helping guide potential model refinements and improvements.

4.6 Final Results

After rigorous training and evaluation, our SER model achieved impressive results on the test dataset. The final results are as follows:

Table II. Performance metrix

Parameters	***Values***
Accuracy	79.94%
Precision	80.36%
Recall	79.28%
F1 Score	79.81%

These results highlight the model's ability to accurately recognize and classify emotions in speech data, demonstrating its Opportunities for practical use in real-world scenarios in understanding and categorizing emotional states from audio inputs.

We can conclude that our experimental setup and results showcase the effectiveness of our SER model and its real-world applicability. The careful data splitting, rigorous model training, and comprehensive evaluation with various classification metrics together underscore the model's performance.

4.7 CONCLUSION

Our Emotion Recognition System (ERS) is based on the innovative "CNN-TM" model. It combines parallel Convolutional Neural Networks (CNNs) and a Transformer encoder to accurately detect emotions in speech signals. The parallel CNN blocks and Transformer encoder jointly enable the model to grasp complex emotional patterns from Mel spectrogram representations. Our supervised training process minimizes cross-entropy loss, leading to precise emotion classification. While our model demonstrates robust performance, future research should explore larger datasets and fine-tuning techniques for further improvements. The "CNN-TM" model holds significant promise for real-time sentiment analysis and human-computer interaction applications.

REFERENCES

[1] L. Wyse, "Audio spectrogram representations for processing with Convolutional Neural Networks," arXiv:1706.09559v1 [cs.SD] ,2017.

[2] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," IEEE, 2017.

[3] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv:1412.6980 [cs.LG], 2015.

[4] V.Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in ICML'10: Proceedings of the 27th International Conference on International Conference on Machine Learning Pages 807-814, 2010.

[5] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova and R. Verma, "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," 10.1109/TAFFC.2014.2336244, 2014.

[6] Y. Gong, Y.-A. Chung and J. Glass, "AST: Audio Spectrogram Transformer," arXiv:2104.01778v3 [cs.SD] , 2021.

[7] A. Vaswani, N. Shazeer, N. Parmar, U. J. L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," arXiv:1706.03762 [cs.CL], 2017.

[8] A. Shukla, K. Vougioukas, P. Ma, S. Petridis and M. Pantic, "Visually Guided Self Supervised Learning of Speech Representations," arXiv:2001.04316 [eess.AS], 2020.

[9] M.-I. Georgescu, R. T. Ionescu, N.-C. Ristea and N. Sebe, "Non-linear Neurons with Human-like Apical Dendrite Activations," arXiv:2003.03229 [cs.NE], 2023.

[10] A. Maas, A. Hannun and A. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," in Proceedings of the 30th International Conference on Machine Learning, Vol. 28, 3., 2013.

[11] A. Gidon, T. Zolnik, P. Fidzinski, F. Bolduan, A. Papoutsi, P. Poirazi, M. Holtkamp, I. Vida and M. Larkum, "Dendritic action potentials and computation in human layer 2/3 cortical neurons," 10.1126/science.aax6239, 2020.

[12] A. Khan, A. Sohail, U. Zahoora and A. S. Qureshi, "A Survey of the Recent Architectures of Deep Convolutional Neural Networks," arXiv:1901.06032 [cs.CV], 2020.

[13] N.-C. Ristea and R. T. Ionescu, "Self-paced ensemble learning for speech and audio classification," arXiv:2103.11988 [cs.SD], 2021.

[14] R. S, A. S. Bharadwaj, D. S.K., M. S. Khadabadi and A. Jayaprakash, "Digital Implementation of the Softmax Activation Function and the Inverse Softmax Function," in IEEE, 2023.