A Major Project Report

On

CNN-BASED EMOTION DETECTION MODEL USING SPEECH RECOGNITION AND FACIAL EXPRESSION FROM IMAGES AND VIDEOS

Submitted to JNTU HYDERABAD

In Partial Fulfilment of the requirements for the Award of Degree of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

Submitted By

M.VASANTH REDDY	**(208R1A1237)**
K.SUDEEPTHI	**(208R1A1231)**
D.DEEPAK JOSHI	**(208R1A1216)**
A.SAI TEJA	**(208R1A1201)**

Under the Esteemed guidance of

Mr. SRINIVAS REDDY B.

Assistant Professor, Department of IT

Department of Information Technology

CMR ENGINEERING COLLEGE
(UGC AUTONOMOUS)

(Accredited by NAAC & NBA, Approved by AICTE NEW DELHI, Affiliated to JNTU, Hyderabad)(Kandlakoya, Medchal Road, R.R. Dist. Hyderabad-501 401)

A.Y. (2023-2024)

CMR ENGINEERING COLLEGE
(UGC AUTONOMOUS)

(Accredited by NAAC & NBA, Approved by AICTE NEW DELHI, Affiliated to JNTU, Hyderabad)(Kandlakoya, Medchal Road, R.R. Dist. Hyderabad-501 401)

Department of Information Technology

CERTIFICATE

This is to certify that the project entitled **"CNN-Based Emotion Detection Model Using Speech Recognition And Facial Expression From Images And Videos"** is a bonafide work carried out by

M.VASANTH REDDY	**(208R1A1237)**
K.SUDEEPTHI	**(208R1A1231)**
D.DEEPAK JOSHI	**(208R1A1216)**
A.SAI TEJA	**(208R1A1201)**

in partial fulfilment of the requirement for the award of the degree of **BACHELOR OF TECHNOLOGY** in **INFORMATION TECHNOLOGY** from CMR Engineering College, affiliated to JNTU, Hyderabad, under our guidance and supervision.

The results presented in this project have been verified and are found to be satisfactory. The results embodied in this project have not been submitted to any other university for the award of any other degree or diploma.

Internal Guide
Mr. SRINIVAS REDDY B
Assistant Professor
Department of IT
CMREC, Hyderabad

Head of the Department
Dr. MADHAVI PINGILI
Professor & HOD
Department of IT
CMREC, Hyderabad

DECLARATION

This is to certify that the work reported in the present project entitled **"CNN-Based Emotion Detection Model Using Speech Recognition And Facial Expression From Images And Videos"** is a record of bonafide work done by us in the Department of Information Technology, CMR Engineering College, JNTU Hyderabad. The reports are based on the project work done entirely by us and not copied from any other source. We submit our project for further development by any interested students who share similar interests to improve the project in the future.

The results embodied in this project report have not been submitted to any other University or Institute for the award of any degree or diploma to the best of our knowledge and belief.

M.VASANTH REDDY (208R1A1237)

K.SUDEEPTHI REDDY (208R1A1231)

D.DEEPAK JOSHI (208R1A1216)

A.SAI TEJA (208R1A1201)

ACKNOWLEDGEMENT

We are extremely grateful to **Dr. A. Srinivasula Reddy**, Principal and **Dr. Madhavi Pingili**, HOD,

Department of IT, CMR Engineering College for their constant support.

We are extremely thankful to **Mr.SRINIVAS REDDY B,** Assistant Professor, Internal Guide, Department of IT, for her constant guidance, encouragement and moral support throughout the project.

We will be failing in duty if we do not acknowledge with grateful thanks to the authors of thereferences and other literatures referred in this Project.

We express our thanks to all staff members and friends for all the help and co-ordination extended in bringing out this project successfully in time.

Finally, We are very much thankful to our parents who guided us for every step.

M.VASANTH REDDY (208R1A1237)

K.SUDEEPTHI REDDY (208R1A1231)

D.DEEPAK JOSHI (208R1A1216)

A.SAI TEJA (208R1A1201)

CONTENTS

ABSTRACT

In recent years, the detection of facial emotions has gained considerable attention as it has diverse applications in fields like psychology, human-computer interaction, and marketing. One of the key challenges in speech emotion recognition is to extract the emotional features effectively from a speech utterance. Despite the promising results of recent studies, they generally do not leverage advanced fusion algorithms for the generation of effective representations of emotional features in speech utterances. To enhance the accuracy of facial emotion detection systems, convolutional neural networks (CNNs) have shown immense potential. This study presents a CNN-based approach for facial emotion detection that uses transfer learning and data augmentation techniques to boost the generalization of the model. The proposed approach was tested on several benchmark datasets, including FER-2013, CK+, and JAFFE databases, and outperformed the existing state-of-the-art models. The results indicate that the CNN- based approach is highly effective in accurately recognizing facial emotions and has significant potential for real-world applications of facial emotion detection. We stack two parallel CNNs for spatial feature representation in parallel to a Transformer encoder for temporal feature representation, thereby simultaneously expanding the filter depth and reducing the feature map with an expressive hierarchical feature representation at a lower computational cost. We use the RAVDESS dataset to recognize eight different speech emotions. We augment and intensify the variations in the dataset to minimize model overfitting. Additive White Gaussian Noise (AWGN) is used to augment the RAVDESS dataset

Keywords: facial emotion, convolutional neural network, FER-2013, CK+, RAVDESS dataset

INTRODUCTION

We propose a facial emotion detection system that utilizes the CNN algorithm to identify the emotions of a person captured on camera. The system processes each frame of a video using a deep learning CNN model trained to categorize frames into accident or non-accident categories. CNNs have demonstrated high accuracy levels of over 95% for image classification tasks, and they require less preparation than previous methods. These algorithms are commonly used in computer vision tasks such as object recognition and image categorization, and they are trained on labeled datasets of faces to learn distinctive features like the positioning of the eyes, nose, and mouth. When the system is introduced to a new face, the CNN analyzes the facial characteristics and generates a unique facial embedding for that individual. The system's database can then compare this embedding with others to determine if there is a match. CNN-based facial recognition systems have found use in diverse fields such as social media, advertising, and security and surveillance. However, ethical and privacy concerns regarding facial recognition technology have been raised, particularly regarding potential biases and misuse of the technology .Because of the potential applications in domains such as psychology, human-computer interaction, marketing, security, and surveillance, facial emotion recognition using CNN has become a popular study issue. Accurately identifying emotions from facial expressions has major implications in a variety of settings, including improving human-robot interaction, recognising driver weariness in vehicle safety, and improving customer experience in retail and advertising. Furthermore, CNN-based facial emotion detection has demonstrated promising results in overcoming obstacles associated with classic emotion recognition systems, such as fluctuations in facial expressions and lighting conditions. Researchers want to attain high accuracy rates in identifying emotions from facial expressions using CNN algorithms, which will lead to the development of useful and efficient real-world applications in a variety of disciplines. Speech emotion recognition (SER) systems classify emotions in speech utterances and are vital in advancing the HCI, healthcare, customer satisfaction, social media analysis, stress monitoring, and intelligent systems. Moreover, SER systems are useful in online tutorials, language translation, intelligent driving, and therapy sessions. In a few situations, humans can be substituted by computer-generated characters with the ability to act naturally and communicate convincingly by expressing human-like emotions. Machines need to interpret the emotions carried by speech utterances. Only with such an ability can a completely expressive dialogue based on joint human–machine trust and understanding be accomplished.

LITERATURE SURVEY

Almadhor, A.; Irfan, R.; Gao, J.; Saleem, N.; Rauf, H.T. ; Kadry, S. (2023) " E2E-DASR: End-to-end deep learning-based dysarthric automaticspeech recognition "

This paper proposes a spatio-temporal dysarthric ASR (DASR) system using Spatial Convolutional Neural Network (SCNN) and Multi-Head Attention Transformer (MHAT) to visually extract the speech features, and DASR learns the shapes of phonemes pronounced by dysarthric individuals.

Shilandari, A.; Marvi, H.; Khosravi, H.; Wang, W. (2022) " Speech emotion recognition using data augmentation method by cycle- generative adversarial networks "

This paper, we present a Cycle-GAN for data augmentation and then test it on SER with two classifier networks. The Cycle-GAN generates samples similar to actual data thereby augmenting the dataset with additional samples for emotion classification.

Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W. (2021) " Survey of deep representation learning for speech emotionrecognition "

This paper presents the first comprehensive survey on the important topic of deep representation learning for SER. We highlight various techniques, related challenges and identify important future areas of research. Our survey bridges the gap in the literature since existing surveys either focus on SER with hand-engineered features or representation learning in the general setting without focusing on SER.

Akçay, M.B.; Ogˇ uz, K. (2020) " Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers "

This paper studies make use of the advances in all fields of computing and technology, making it necessary to have an update on the current methodologies and techniques that make SER possible. We have identified and discussed distinct areas of SER, provided a detailed survey of current literature of each, and also listed the current challenges.

Zhang H, Huang B, Tian G. (2020) " Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture "

This paper combines the time and texture information of image sequence to improve the recognition rate. At the decision level, the features of the two channels are fused, and good recognition results are obtained. In this paper, facial region of interest is extracted for expression recognition.

Magdin, M.; Sulka, T.; Tomanová, J.; Vozár, M. (2019) " Voice analysis using PRAAT software and classification of user emotional state "

This paper deals with the user's emotional state classification based on the voice track analysis, it describes its own solution - the measurement and the selection process of appropriate voice characteristics using

ANOVA analysis and the use of PRAAT software for many voice aspects analysis and for the implementation of own application to classify the user's emotional state from his/her voice. In the paper are presented the results of the created application testing and the possibilities of further expansion and improvement of this solution.

He J, Li D, Bo S, Yu L. (2019) " Facial action unit detection with multilayer fused multi-task and multi-label deep learning network "

This paper likely evaluates the proposed model using standard performance metrics on benchmark datasets, providing insights into its effectiveness and potential contributions to the field of facial expression analysis. For more details, accessing the specific paper directly or consulting academic databases is recommended.

Liu, Z.T.; Xie, Q.; Wu, M.; Cao, W.H.; Mei, Y.; Mao, J.W. (2018) " Speech emotion recognition based on an improved brain emotion learningmodel "

This paper, CASIA , SAVEE emotion corpus, and FAU Aibo dataset are used for experiment of speech emotion recognition, in which Low Level Descriptors (LLDs) of MFCC related features and their 1st order delta coefficients are extracted, and the proposal is tested on INTERSPEECH 2009 standard feature set as well to verify the robustness. Meanwhile, LDA, PCA, and PCA+LDA are used to reduce the dimension of the feature set, respectively. In addition, two feature sets of row LLDs and row LLDs+1Δ coefficients are compared in terms of the speech emotion recognition accuracy. In this work, the performance of original BEL model and some conventional methods are also presented for comparison to verify the feasibility of the proposal.

Patel, P.; Chaudhari, A.; Kale, R.; Pund, M. (2017) " Emotion recognition from speech with gaussian mixture models via boosted gmm "

In this paper, the selection of language sentence for experiment analysis mainly comes from twoaspects followed. First, statements selected must not contain a particular aspect of emotionaltendency; secondly, statements selected must contain high emotional freedom, for the samestatement can exert all kinds of emotions. Moreover, to the length of the statement, compositionof consonants and auxiliary components, all differences between male and female should be considered.

Mira J, ByoungChul K, JaeYeal N. (2016) " Facial landmark detection based on an ensemble of local weighted regressors during real driving situation "

This paper proposes a novel facial landmark detection (FLD) algorithm for use in real driving situations. The proposed algorithm is based on an ensemble of local weighted random forest regressor (WRFR) with random sampling consensus (RANSAC) and explicit global shape models, and considers the dynamic and irregular characteristics of driving.

EXISTING SYSTEM

Over the past years, the automatic process of facial emotion recognition (FER) has become a substantial area of interest for researchers. The main goals for FER systems are the identification of a person's emotions and their intensities, followed by the classification of expression cause, which can be genuine or simulated.

From the implementation perspective, in the last years, FER systems developed using different types of artificial neural networks (ANNs), which proved to have better results than using traditional machine learning methods based on feature descriptors such as histogram of oriented gradients (HOG), or local binary pattern (LBP) combined with data classifiers such as support vector machine (SVM), k-nearest neighbors (KNN) or random forest. As demonstrated in other detection or recognition processes based on ANNs, people's emotions can also be accurately detected and recognized in a subject-independent way by building a model through the analysis of a collection of training data from different individuals, including skeletal movements. The use of ANNs for emotion detection and recognition opened many opportunities for practical applications, especially in fields such as healthcare, security, business, education, or manufacturing.

DIS-ADVANTAGES:

- **Data Privacy Concerns:** Collecting and processing facial expressions and speech data raises privacy concerns. Users may be uncomfortable with the idea of their emotions being analyzed, particularly if the data is not handled securely.

- **Data Bias:** The model's accuracy can be affected by biases in the training data. If the dataset used for training is not diverse enough or is biased towards certain demographics, the model may not generalize well to different populations.

- **Multimodal Integration Challenges:** Integrating information from different modalities (speech and facial expressions) can be complex. Aligning data from these sources and creating a cohesive representation for the model can be challenging.

- **Real-time Processing Requirements:** Achieving real-time processing for both speech and facial expressions in videos can be computationally demanding. This may limit the practicality of deploying the model in real-time applications, especially on devices with limited processing power.

- **Ambiguity in Emotion Recognition:** Emotions are complex and can vary across individuals and cultures. There may be instances where the model struggles to accurately interpret ambiguous or subtle emotional cues, leading to misclassifications.

PROPOSED SYSTEM

The System we proposed is facial emotion using CNN Algorithm. The goal is to develop a systemthat uses a camera to identify a person's emotions. The goal is to process every frame of a video via a deep learning convolution neural network model that has been trained to categorize video frames into accident- or non-accident-related categories. Convolutional Neural Networks have shown to be a quick and reliable method for classifying photos. In comparison to previous image classification methods, CNN-based image classifiers have achieved accuracy levels of over 95% for relatively smaller datasets. They also require less preparation. Computer vision tasks like object recognition and image categorization frequently employ CNNs, a kind of deep learning algorithm. CNNs are trained on a sizable dataset of labelled faces for facial recognition in order to learn the distinctive features and traits of various faces, such as the placement of the eyes, nose, and mouth. When a fresh face is introduced to the system, CNN examines the facial traits and creates a distinctive image of the face known as a face embedding. The database of the system's system can then be used to compare this embedding to the embeddings of other faces to see if there is a match. Recent years have seen a rise in the use of CNN facial recognition systems in a variety of fields, including social media, advertising, and security and surveillance. Yet there are also worries about the ethical and privacy ramifications of facial recognition technology, especially in terms of potential biases and the danger of abuse.

ADVANTAGES:

- **Define the Problem:** Clearly define the problem you are addressing, such as real-time emotion detection from multimedia sources (images and videos) using both speech and facial expressions.
- **Preprocessing:** Preprocess both image and audio data:

  Image preprocessing: Resize images, normalize pixel values, and augment data if needed.

  Audio preprocessing: Extract features (e.g., MFCCs), normalize, and consider data augmentation.
- **Model Architecture:** Design a CNN-based model for facial expression recognition and a separate model for speech emotion recognition. Here's a high-level overview:

  Facial Expression Model:

  Use a pre-trained CNN for image recognition or train a CNN for facial expression analysis.

  Incorporate techniques like transfer learning for better performance.

  Speech Emotion Model:

  Use a pre-trained model for speech emotion recognition (or train a model on a speech emotion dataset).

  Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) can be effective for speech analysis.

IMPLEMENTATION

Step 1: Set Up Your Environment

Install Required Libraries:

- TensorFlow or PyTorch for deep learning.
- OpenCV for image and video processing.
- Speech recognition library (e.g., SpeechRecognition) for processing audio.
- Other necessary libraries like NumPy, Matplotlib, etc.

Step 2: Data Collection and Preprocessing

Dataset:

- Collect or use existing datasets for facial expression recognition and emotion detection.
- Consider datasets with audio recordings for speech-based emotion analysis.

Data Preprocessing:

- Preprocess facial images (resize, normalize, etc.).
- Extract facial landmarks or use pre-trained facial landmark models.
- Preprocess audio data (spectrogram, MFCC extraction, etc.).

Step 3: Build the Facial Expression Recognition Model

Facial Expression Model Architecture:

- Use a pre-trained CNN model (e.g., VGG16, ResNet) for image-based facial expression recognition.
- Fine-tune the pre-trained model on your emotion dataset.

Step 4: Build the Speech Emotion Recognition Model

Speech Emotion Model Architecture:

- Convert audio data into spectrograms or use MFCCs.
- Build a CNN or LSTM-based model for speech emotion recognition.
- Train the model on a speech emotion dataset.

Step 5: Combine Models

Integration:

- Decide how to integrate the facial expression and speech emotion models.
- Concatenate their respective feature vectors before feeding them to a fusion model.

Step 6: Fusion Model

Fusion Architecture:

- Design a neural network that combines the features from both the facial expression and speech emotion models.
- Train the fusion model on a combined dataset.

Step 7: Testing and Evaluation

Testing:

- Evaluate your model on a separate test set that includes both facial expression and speech samples.

Metrics:

- Use appropriate evaluation metrics (accuracy, F1 score, confusion matrix) for both facial expression and speech emotion recognition.

Step 8: Deployment

Deployment:

- Deploy your model using frameworks like TensorFlow Serving, Flask, or FastAPI.

Integration:

- Integrate the model into your application for real-time or batch processing.

Step 9: Fine-Tuning and Optimization

Optimization:

- Fine-tune your models based on performance feedback.
- Optimize for real-time processing if required.

Step 10: Maintenance

Maintenance:

- Regularly update and maintain your model to ensure it performs well over time.

CONVOLUTIONAL NEURAL NETWORK(CNN):

One of the artificial neural network, used in image recognitionis CNN.

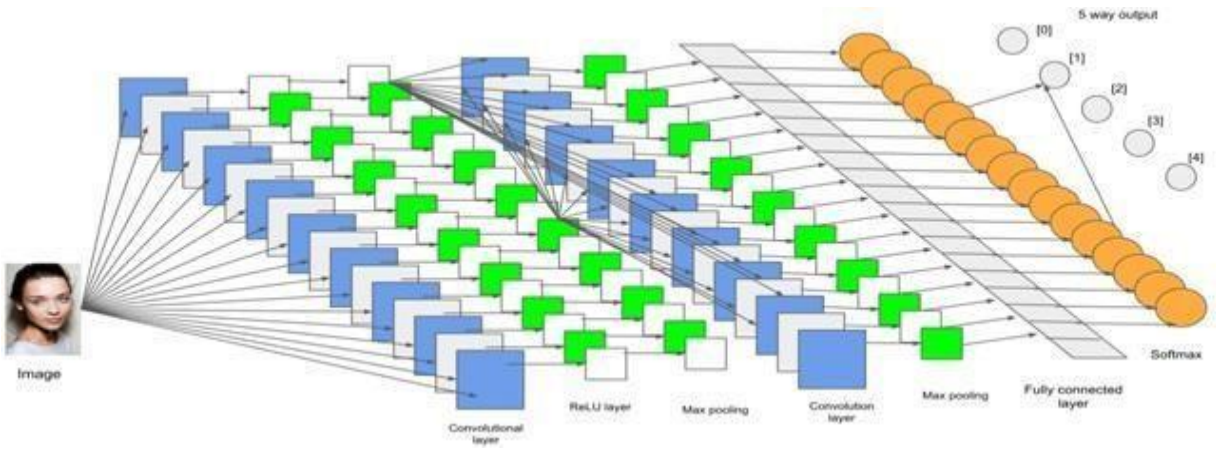
A CNN is a type of artificial neural network used in image recognition. This comes under category of multilayer perceptron.

The are four layers in CNN :

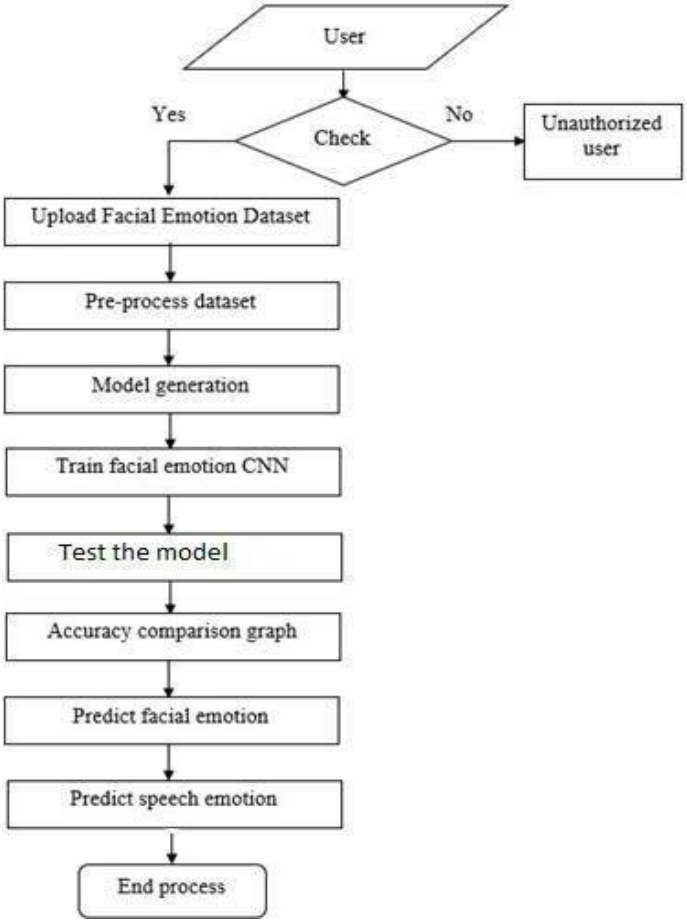
- Input layer
- Output layer
- Hidden layer
- Normalization layer

The Network Layers:

- **Image Input Layer** An Image Input Layer is where you specify the image size
- **Convolutional Layer** It is a CNN filter , where inputs are filter size and number of neurons.
- **Batch Normalization Layer** Batch normalization layers normalize the activations and gradients propagating through a network.
- **ReLU Layer** It is a linear rectified unit , it is used to convert negative feature to 0.
- **Max-Pooling Layer** It is used for down sampling and toreduce redundant features.
- **Fully Connected Layer** It is used to connect all neuronsand we provide number of classes in it.
- **Softmax Layer** It is used to find out the probability ofobject in the image.
-

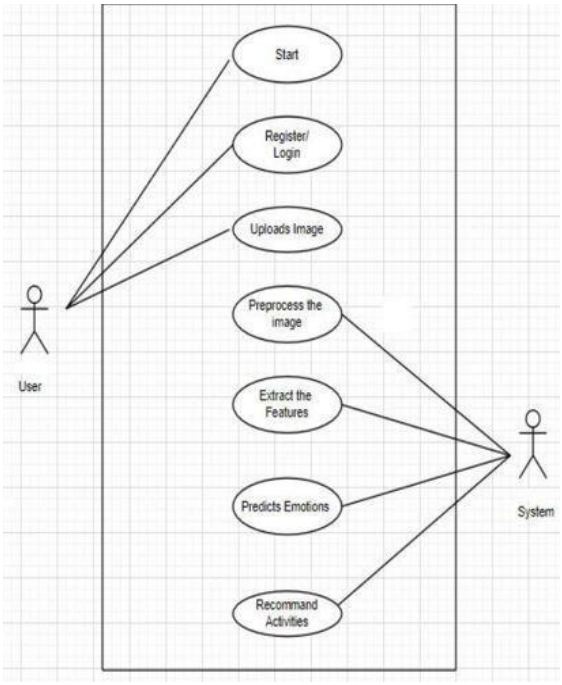


SYSTEM ARCHITECTURE

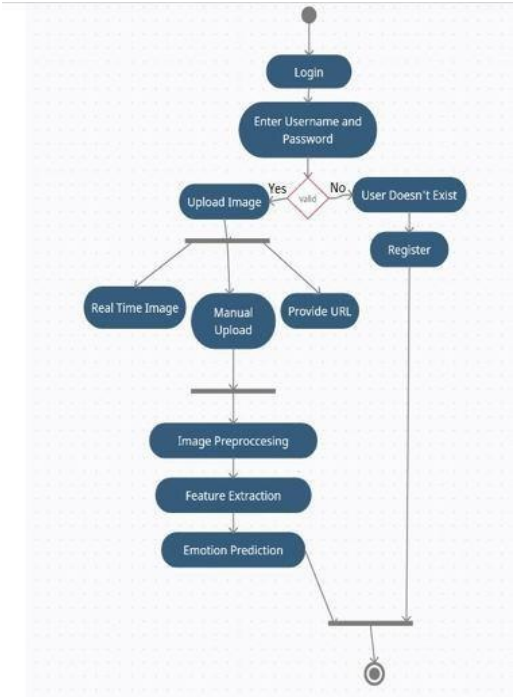


UML DIAGRAMS

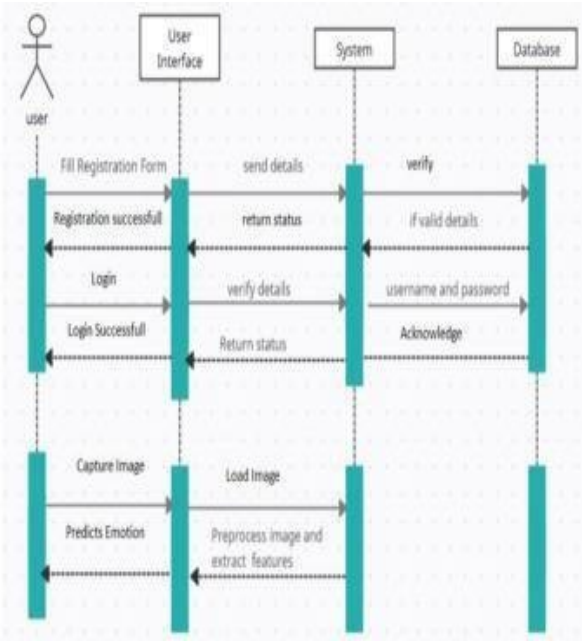
a) Use Case Diagram



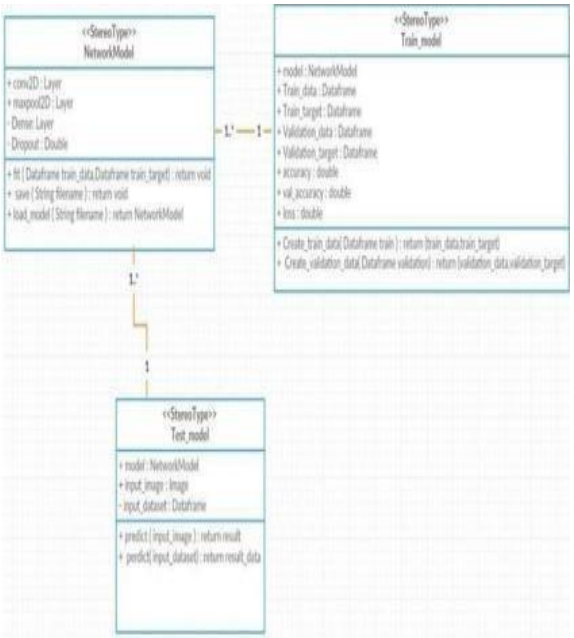
b) Activity Diagram



c) Sequence Diagram



d) Class Diagram



REFERENCES

[1]Almadhor, A.; Irfan, R.; Gao, J.; Saleem, N.; Rauf, H.T. ; Kadry, S., "E2E-DASR: End-to-end deep learning-based dysarthric automaticspeech recognition", 2023.

[2]Shilandari, A.; Marvi, H.; Khosravi, H.; Wang, W., "Speech emotion recognition using data augmentation method by cycle- generative adversarial networks", 2022.

[3]Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B.W., "Survey of deep representation learning for speech emotionrecognition", 2021.

[4]Akçay, M.B.; Ogˇ uz, K., "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers", 2020

[5]Zhang H, Huang B, Tian G., "Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture", 2020.

[6]Magdin, M.; Sulka, T.; Tomanová, J.; Vozár, M., "Voice analysis using PRAAT software and classification of user emotional state", 2019.

[7]He J, Li D, Bo S, Yu L., "Facial action unit detection with multilayer fused multi-task and multi-label deep learning network", 2019.

[8]Hossain MS, Muhammad G., "Emotion recognition using deep learning approach from audio–visual emotional big data",2019

[9] Liu, Z.T.; Xie, Q.; Wu, M.; Cao, W.H.; Mei, Y.; Mao, J.W., "Speech emotion recognition based on an improved brain emotion learningmodel", 2018.

[10] Patel, P.; Chaudhari, A.; Kale, R.; Pund, M., "Emotion recognition from speech with gaussian mixture models via boosted gmm", 2017.

[11] Ghimire D, Jeong S, Lee J, Park SH., "Facial expression recognition based on local region specific features and support vector machines",2017

[12] Mira J, ByoungChul K, JaeYeal N., "Facial landmark detection based on an ensemble of local weighted regressors during real driving situation", 2016.

[13] Zhang T, Zheng W, Cui Z, Zong Y, Yan J, Yan K., " A deep neural network-driven feature learning method for multi-view facial expression recognition", 2016

[14] Jung H, Lee S, Yim J, Park S, Kim J., "Joint fine-tuning in deep neural networks for facial expression recognition", 2016