A Project Report On

DYNAMIC EMOTION RECOGNITION USING HYBRID DEEP LEARNING MODELS

Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

Submitted By

TATA P N V S SATYANARAYANA MURTHY	**20P31A1258**
NEKKANTI SAI CHAITANYA	**20P31A1239**
KARRI MOUSAMI REDDY	**20P31A1221**
BIRADHA DILEEP KRISHNA KUMAR	**20P31A1208**

Under the esteemed supervision of

Mrs. N. SURYA KALA., M.Tech.,

Assistant Professor

DEPARTMENT OF INFORMATION TECHNOLOGY

ADITYA COLLEGE OF ENGINEERING & TECHNOLOGY (A)

Permanently Affiliated to JNTUK, Kakinada * Approved by AICTE New Delhi

Accredited by NBA, Accredited by NAAC (A+) with 3.4 CGPA

Aditya Nagar, ADB Road, Surampalem, Kakinada District, Andhra Pradesh.

2020-2024

ADITYA COLLEGE OF ENGINEERING & TECHNOLOGY(A)

(An Autonomous Institution)

Permanently Affiliated to JNTUK, Kakinada * Approved by AICTE New Delhi
Accredited by NBA, Accredited by NAAC (A+) with 3.4 CGPA
Aditya Nagar, ADB Road, Surampalem, Kakinada District, Andhra Pradesh

DEPARTMENT OF INFORMATION TECHNOLOGY

CERTIFICATE

This is to certify that the project work entitled "**DYNAMIC EMOTION RECOGNITION USING HYBRID DEEP LEARNING MODELS** ", is a bonafide work carried out by **TATA P N V S SATYANARAYANA MURTHY (20P31A1258), NEKKANTI SAI CHAITANYA (20P31A1239), KARRI MOUSAMI REDDY (20P31A1221), BIRADHA DILEEP KRSIHNA KUMAR (20P31A1208)**, in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology** from **Aditya College of Engineering & Technology** during the academic year 2020-2024.

Project Guide
Mrs. N. Surya Kala, M.Tech.,
Assistant Professor

Head Of The Department
Mr. R V V N Bheema Rao, M.Tech., (Ph.D.)
Associate Professor

EXTERNAL EXAMINER

DECLARATION

We hereby declare that this project entitled "**Dynamic Emotion Recognition Using Hybrid Deep Learning Models**", has been undertaken by us and this work has been submitted to **Aditya College of Engineering & Technology** affiliated to JNTUK, Kakinada, in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology**.

We further declare that this project work has not been submitted in full or part for the award of any degree of this or in any other educational institutions.

PROJECT ASSOCIATES

TATA P N V S SATYANARAYANA MURTHY	**20P31A1258**
NEKKANTI SAI CHAITANYA	**20P31A1239**
KARRI MOUSAMI REDDY	**20P31A1221**
BIRADHA DILEEP KRISHNA KUMAR	**20P31A1208**

ACKNOWLEDGEMENT

It is with immense pleasure that we would like to express our indebted gratitude to our Project Supervisor, **Mrs. SuryaKala., M.Tech.,** who has guided us a lot and encouraged us in every step of the project work, her valuable moral support and guidance throughout the project helped us to a great extent.

We wish to express our sincere thanks to the Head of the Department **Mr. R V V N Bheema Rao M.Tech., (Ph. D)** for his valuable guidance given to us throughout the period of the project work and throughout the program.

We feel elated to thank **Dr. Ch V Raghavendran Ph.D** Dean – Academics of Aditya College of Engineering & Technology for his cooperation in completion of our project and throughout the program.

We feel elated to thank **Dr. D Kishore Ph.D** Dean – Evaluation of Aditya College of Engineering & Technology for his cooperation in completion of our project and throughout the program.

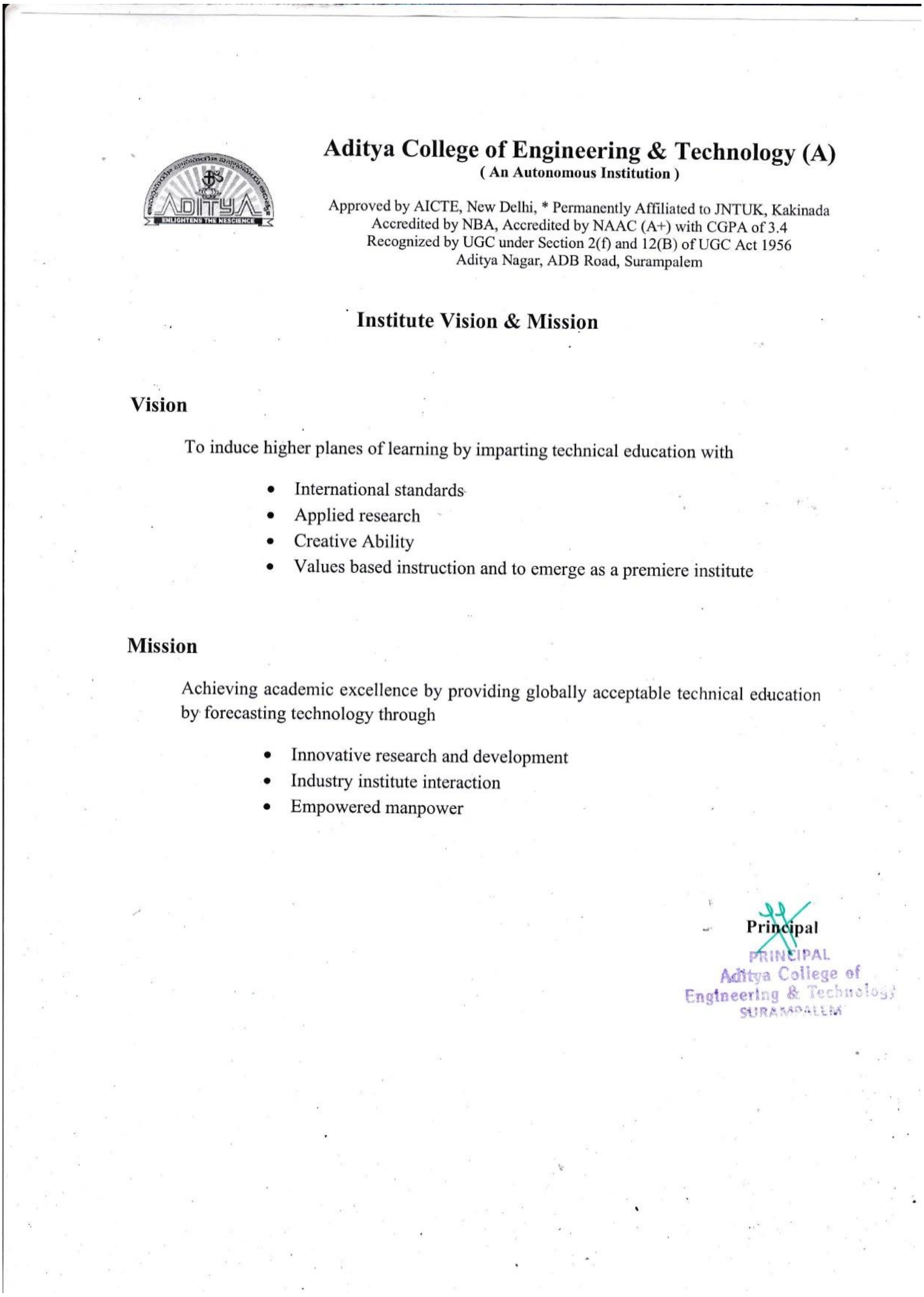
We feel elated to thank **Dr. Dola Sanjay S Ph.D** Principal of Aditya College of Engineering & Technology for his cooperation in completion of our project and throughout the program.

We wish to express our sincere thanks to all **faculty members, lab programmers** for their valuable assistance throughout the period of the project.

We avail this opportunity to express our deep sense and heart full thanksto the Management of **Aditya College Of Engineering & Technology** for providing a great support for us in completing our project and also throughout the program.

PROJECT ASSOCIATES

TATA P N V S SATYANARAYANA MURTHY	**20P31A1258**
NEKKANTI SAI CHAITANYA	**20P31A1239**
KARRI MOUSAMI REDDY	**20P31A1221**
BIRADHA DILEEP KRISHNA KUMAR	**20P31A1208**



Aditya College of Engineering & Technology (A)

(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Institute Vision & Mission

Vision

To induce higher planes of learning by imparting technical education with

- International standards
- Applied research
- Creative Ability
- Values based instruction and to emerge as a premiere institute

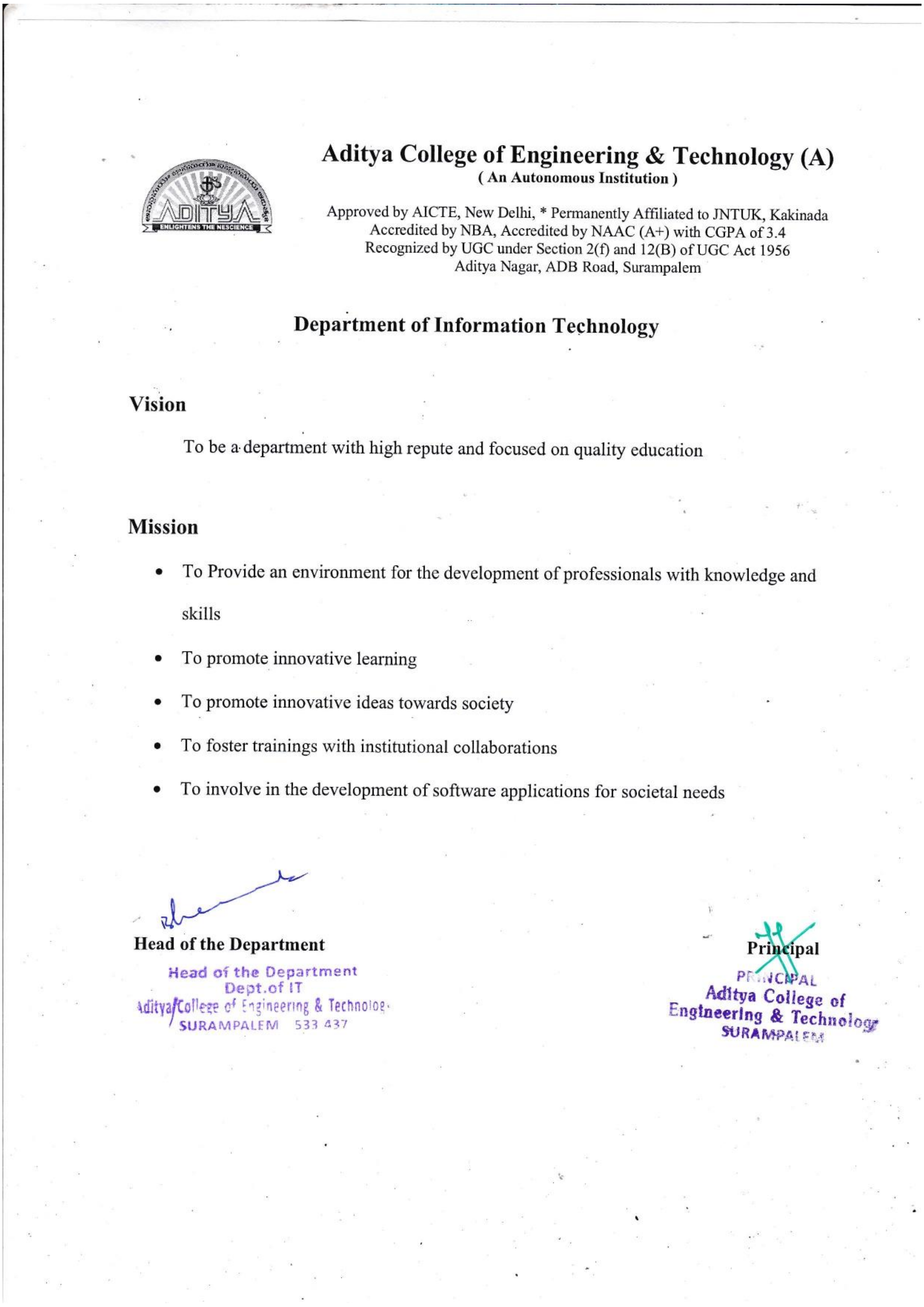
Mission

Achieving academic excellence by providing globally acceptable technical education by forecasting technology through

- Innovative research and development
- Industry institute interaction
- Empowered manpower

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM



Aditya College of Engineering & Technology (A)

(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Vision

To be a department with high repute and focused on quality education

Mission

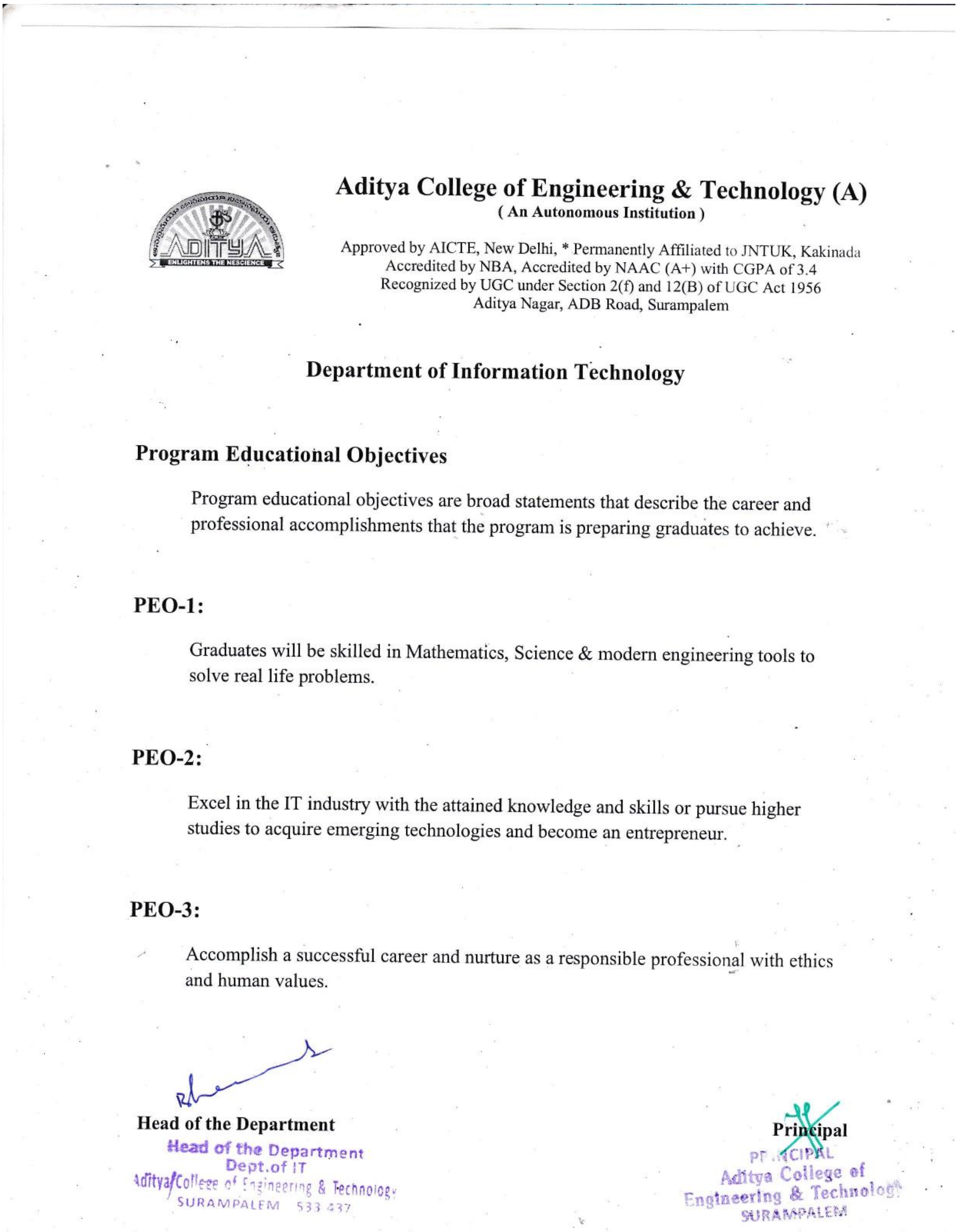
- To Provide an environment for the development of professionals with knowledge and skills
- To promote innovative learning
- To promote innovative ideas towards society
- To foster trainings with institutional collaborations
- To involve in the development of software applications for societal needs

Head of the Department

Head of the Department
Dept.of IT
Aditya College of Engineering & Technology
SURAMPALEM 533 437

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM



Aditya College of Engineering & Technology (A)
(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Program Educational Objectives

Program educational objectives are broad statements that describe the career and professional accomplishments that the program is preparing graduates to achieve.

PEO-1:

Graduates will be skilled in Mathematics, Science & modern engineering tools to solve real life problems.

PEO-2:

Excel in the IT industry with the attained knowledge and skills or pursue higher studies to acquire emerging technologies and become an entrepreneur.

PEO-3:

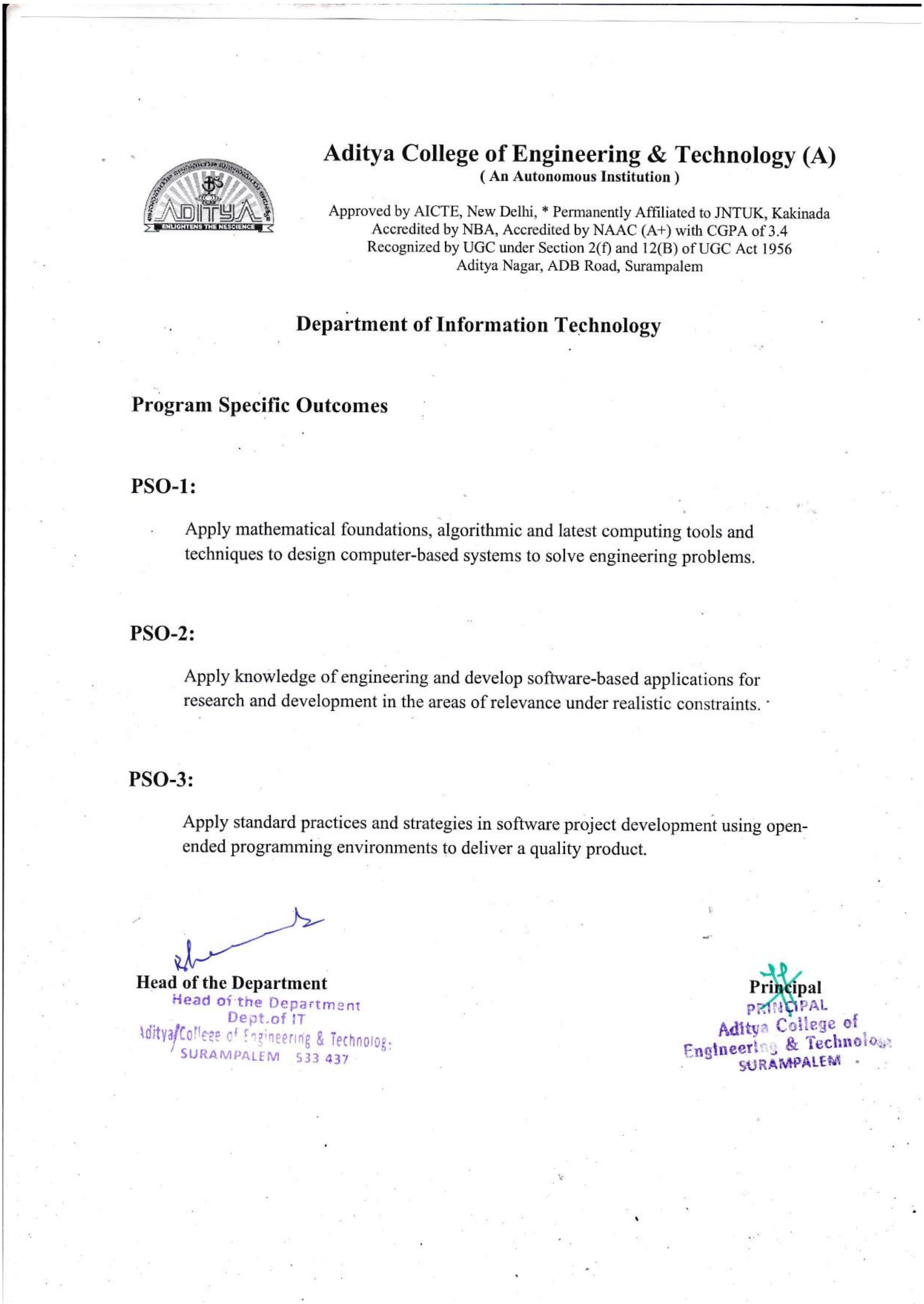
Accomplish a successful career and nurture as a responsible professional with ethics and human values.

Head of the Department

Head of the Department
Dept.of IT
Aditya College of Engineering & Technology
SURAMPALEM 533 437

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM



Aditya College of Engineering & Technology (A)
(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Program Specific Outcomes

PSO-1:

Apply mathematical foundations, algorithmic and latest computing tools and techniques to design computer-based systems to solve engineering problems.

PSO-2:

Apply knowledge of engineering and develop software-based applications for research and development in the areas of relevance under realistic constraints.

PSO-3:

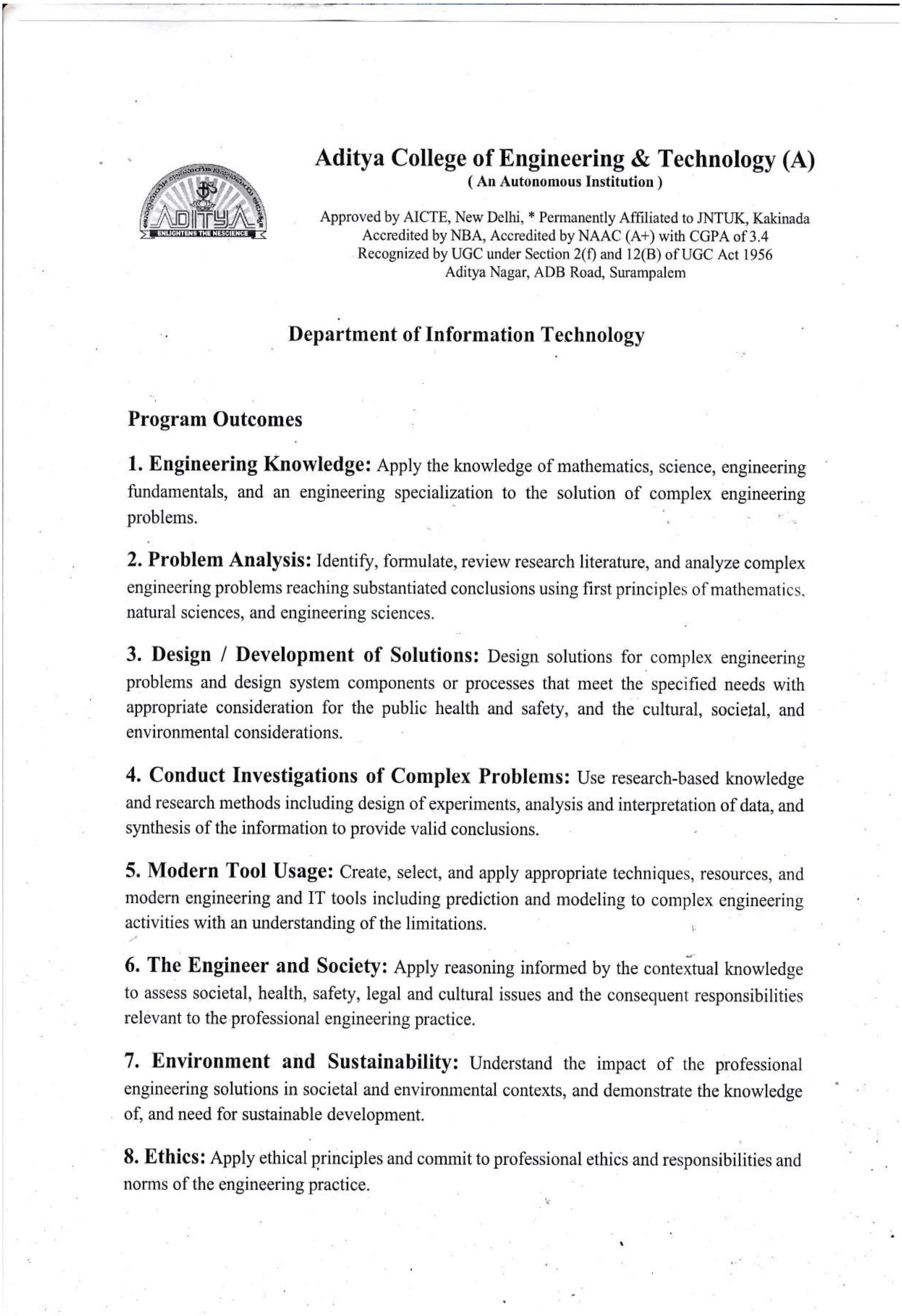
Apply standard practices and strategies in software project development using open-ended programming environments to deliver a quality product.

Head of the Department

Head of the Department
Dept.of IT
Aditya College of Engineering & Technology
SURAMPALEM 533 437

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM



Aditya College of Engineering & Technology (A)
(An Autonomous Institution)

Approved by AICTE, New Delhi, * Permanently Affiliated to JNTUK, Kakinada
Accredited by NBA, Accredited by NAAC (A+) with CGPA of 3.4
Recognized by UGC under Section 2(f) and 12(B) of UGC Act 1956
Aditya Nagar, ADB Road, Surampalem

Department of Information Technology

Program Outcomes

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem Analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design / Development of Solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

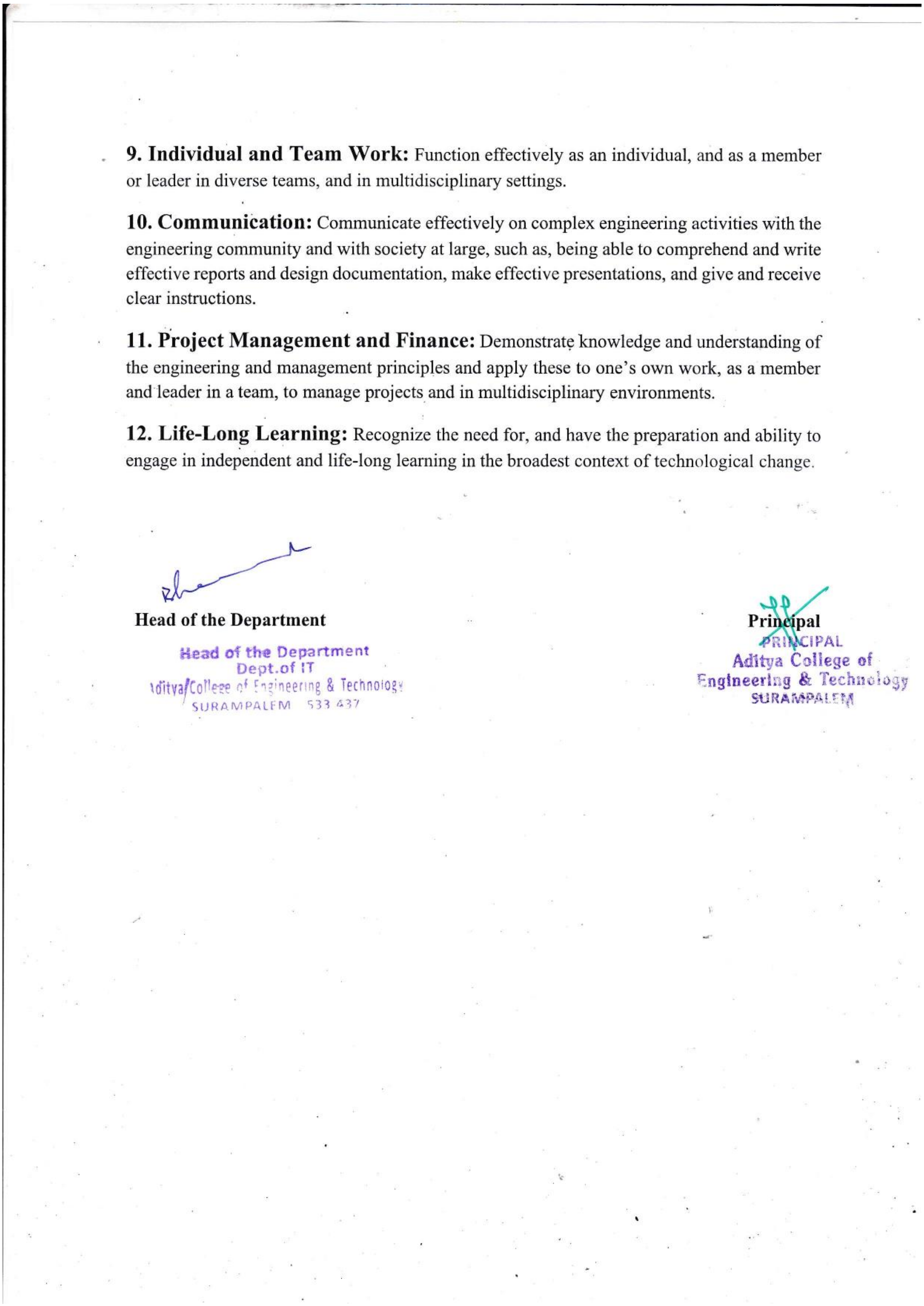
4. Conduct Investigations of Complex Problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The Engineer and Society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and Sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.



9. Individual and Team Work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. Project Management and Finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. Life-Long Learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Head of the Department

Head of the Department
Dept.of IT
Aditya College of Engineering & Technology
SURAMPALEM 533 437

Principal

PRINCIPAL
Aditya College of
Engineering & Technology
SURAMPALEM

ABSTRACT

Emotion recognition is an advanced technology used to interpret and classify human emotions based on facial expressions and non-verbal cues. According to the history involved, spoken communication reflects 7% of the message in personal communication. Voice gives 38% and facial expression produces 55% of the message. As per research, the prominent descriptors of feelings of humans are “Emotions’’. It is essential for simple and straightforward recognition of human psychology at particular instant without really asking them. Users need human-like interactions to better communicate with computers i.e., Human Computer Interaction (HCI) which is essential for the humanization of AI.

Facial expressions are a fundamental aspect of human communication that convey emotions, intentions, and reactions. Facial emotion recognition (FER) is a rapidly evolving field in artificial intelligence that aims to automatically identify emotions from facial expressions. It leverages computer vision techniques to analyze facial features like eyes, eyebrows, and mouth movements. By extracting these visual cues, FER systems can classify emotions such as happiness, anger, sadness, surprise and fear. This technology has numerous potential applications, including human-computer interaction systems that adapt to user emotions, educational tools that gauge student engagement, and market research studies that analyze emotional responses to stimuli. However, challenges remain in ensuring robust performance across diverse populations and capturing the nuances of human emotions. As research progresses, FER holds promise for revolutionizing how we interact with machines and understand human emotions in various contexts. In this study, we are developing a model for emotion recognition using a hybrid neural network combination of Convolutional Neural Networks (CNN) and Long-Short-Term-Memory (LSTM). LSTM will be an added advantage as it recognises the pattern in the person’s emotion while detection.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

INTRODUCTION

CHAPTER - 1

INTRODUCTION

1.1 Introduction

In recent years, the need to recognize a person’s emotions has increased, and there has been a growing interest in human emotion recognition across various fields, including brain–computer interfaces, assistance, medicine, psychology, and marketing. Facial expressions are one of the primary nonverbal means of conveying emotion and play an important role in everyday human communication.

Emotion detection, also known as Facial Emotion Recognition (commonly known as FER), is a fascinating field within the realm of artificial intelligence and computer vision. It involves the identification and interpretation of human emotions from facial expressions. It is one of the most researched fields of computer vision till date and is still in continuous evaluation and improvement.

Facial expressions not only play a vital role in daily life communication like showing anger, happiness and sorrow but they also provide a lot of hidden and non-verbal information for important tasks like video surveillance and monitoring, customer feedback analysis, mental health monitoring and last but not the least human robot interaction / Human Computer Interaction (HCI). Emotions are generally classified as Happy, Surprise, Neutral, Fear, Sad, Anger.

Convolutional Neural Networks (CNNs) have achieved remarkable success in image recognition tasks. CNNs excel at extracting spatial features from individual video frames. However, emotions often unfold over time, with subtle changes in facial expressions. A single frame might not capture the entire emotion. CNNs typically process video frames independently, neglecting the temporal relationships between frames.

Temporal dependencies refer to the relationships and connections between information across different points in time. In simpler terms, it's the "how things change over time" aspect of your data. For example, in video emotion recognition, imagine you see a person's mouth turn down slightly, followed by furrowed brows and then tears welling up. These individual expressions, spread across different frames, are connected and tell a story of sadness. Capturing these relationships is crucial for understanding the overall emotional state.

LSTMs (Long Short-Term Memory) networks are specifically designed to handle temporal dependencies. They have internal mechanisms that remember past information and use it to understand the current context, making them ideal for analyzing sequences of frames.

1.2 Literature Survey

Facial Emotion Recognition (FER) has seen significant advancements with Convolutional Neural Networks (CNNs), effectively categorizing emotions into seven groups. CNNs excel in image processing, maintaining model concepts while reducing boundaries. Techniques like dropout combat overfitting, ensuring model robustness. FER is crucial for human-computer interaction, with architectures like VGGNet enhancing accuracy.

Integration of datasets like FER 2013, CK, and KDEF diversifies facial expressions, improving system robustness. Strategies like data augmentation and transfer learning further enhance FER performance. Challenges remain, including facial occlusions and lighting variations, requiring innovative solutions and interdisciplinary collaboration.

The future holds promise for FER, driven by deep learning and multimodal fusion, unlocking new possibilities in affective computing. Interdisciplinary efforts will advance understanding of human emotions, fostering empathetic computational systems [1].

The paper delves into emotion recognition employing CNNs, offering multiple classification models. It scrutinizes the balance between performance and inference latency in real-time scenarios. Diverse data preprocessing techniques for FER2013 dataset are proposed to tackle inherent challenges.

FER2013 comprises 35,887 grayscale images sorted into seven emotion categories. The research juxtaposes multi-class and binary classification methodologies, emphasizing the advantages of distinct classifiers for heightened accuracy [2].

The paper amalgamates online learning and machine learning methodologies for monitoring student emotions within educational settings. It employs machine-learning models to forecast student emotions and engagement in online courses, prioritizing real-time emotional feedback.

Ensemble methods are underscored for their enhanced accuracy and resilience in classification duties, particularly with multiclass datasets such as facial expression recognition. The significance of face detection techniques is emphasized for extracting facial landmarks and emotions within FER datasets, ultimately amplifying the overall classification efficacy [3].

The paper concentrates on amalgamating the Haar cascade classifier and CNN model to achieve real-time facial expression recognition. It leverages data augmentation methods such as image flipping to enrich facial image datasets, thereby augmenting model robustness and generalization. Haar-like features serve as pivotal components for extracting visual features within the model architecture, facilitating efficient representation of facial expressions.

By integrating the Haar cascade classifier with CNNs, the paper aims to capitalize on the complementary strengths of both approaches, combining the speed and simplicity of Haar cascades with the discriminative power of CNNs for accurate and efficient facial expression recognition in real-time scenarios. Data augmentation techniques, particularly image flipping, contribute to the diversification and enrichment of the training dataset, mitigating overfitting and enhancing model performance across various facial expressions and poses.

This integrative approach holds promise for advancing the state-of-the-art in real-time facial expression recognition systems, with potential applications in affective computing, human-computer interaction, and emotion-aware technologies [4].

The paper introduces a novel approach for facial expression recognition in videos, combining Recurrent Neural Networks (RNN) with Convolutional Neural Networks (CNN). This fusion achieves a test accuracy of 61% on the RAVDESS dataset.

The methodology entails feeding the CNN output into a feedforward pass up to an intermediate fully connected layer to extract emotion features effectively. Furthermore, the traditional 2D CNNs are extended to incorporate 3D CNNs, enabling the model to capture spatio-temporal patterns in video data, thereby enhancing its performance in facial expression recognition tasks [5].

Facial expressions serve as vital cues for non-verbal communication in the digital realm, augmenting human-computer interactions by recognizing situation-dependent emotions. Past research has explored diverse methodologies including Bayesian Networks, Neural Networks, and Hidden Markov Models for facial expression recognition. However, recent focus has shifted towards deep learning, particularly Convolutional Neural Networks (CNN), due to their promising results in feature extraction and expression classification. In this study, 48x48 grayscale images from Kaggle's ICMP 2013-FER dataset are utilized for training CNN models, showcasing the efficacy of deep learning approaches in facial emotion recognition tasks [6].

Research in the domain of deep learning has extensively explored facial emotion recognition, hand gesture recognition, and speech conversion. Recent studies have investigated the use of deep convolutional generative adversarial networks (DCGAN) for training facial expressions and gestures, showcasing their potential in capturing complex patterns and generating realistic outputs. Furthermore, deep learning models have demonstrated high identification rates in recognizing gestures and facial expressions, underscoring their effectiveness in diverse recognition tasks. Additionally, hybrid models such as Long Short-Term Memory (LSTM) networks have been employed for training emotional speech recognition systems, leveraging their ability to capture temporal dependencies and nuances in speech data [7].

The paper conducts a comprehensive review of pertinent literature concerning metric learning and Siamese Neural Networks. It introduces deep Siamese Neural Networks tailored for facial expression recognition, particularly emphasizing verification and identification tasks. The algorithm's efficacy is assessed across multiple datasets including AffectNet, FER2013, and Compound Facial Expressions of Emotion (CFEE), where it demonstrates superior performance compared to other deep learning-based approaches. Furthermore, the research delves into models aimed at reducing intra-class variation and augmenting inter-class variation, culminating in the exploration of ensemble methods to bolster performance across diverse datasets [8].

The paper provides a literature survey focusing on CLCM, a lightweight CNN model designed for facial emotion recognition. Despite its compact architecture, CLCM has demonstrated superior performance compared to several existing models, making it well-suited for deployment on mobile devices and real-time emotion-based studies.

In comparative evaluations, CLCM exhibited competitive performance against well-known models such as Inception V3, EfficientNet-B0, ResNet-50, and DenseNet121 across multiple datasets. The primary aim of CLCM is to optimize performance for devices with limited computational capacity, addressing the growing demand for efficient and practical solutions in facial emotion recognition applications

[9].

Facial emotion recognition systems encompass diverse methodologies aimed at extracting valuable insights from facial data. Machine learning algorithms, particularly deep learning techniques, have emerged as frontrunners in accurately classifying emotions, owing to their ability to decipher nuanced facial expressions realistically. Certain systems concentrate on scrutinizing specific facial regions such as the eyes and mouth for emotion recognition, achieving commendable success rates averaging around 70 percent. This targeted approach underscores the significance of localized facial cues in emotion identification. Research underscores the pivotal role of facial expressions in non-verbal communication and human-computer interaction. Facial emotion recognition systems play a crucial role in augmenting these interactions by enabling machines to interpret and respond to human emotions effectively, fostering more intuitive and empathetic communication channels [10].

The paper provides a literature survey centered on facial expression recognition employing deep learning, with a particular emphasis on a Convolutional Neural Network (CNN) based on the LeNet architecture. The model's objective is to classify a range of facial expressions including fear, anger, happiness, surprise, sadness, neutrality, and disgust, achieving an accuracy rate of 60.37%. Challenges associated with accurately identifying facial expressions are discussed, highlighting the complexities and variations inherent in facial expressions. Future research directions include expanding the model to process color images and exploring the utilization of pre-trained models such as AlexNet or VGGNet for improved facial emotion recognition performance [11].

The paper provides a literature survey focusing on enhancing Facial Emotion Recognition (FER) through Transfer Learning (TL) utilizing Deep Convolutional Neural Network (DCNN) models. Researchers leveraged pretrained DCNN models including Inception-v3, VGG-19, ResNet-152, ResNet-50, and DenseNet-161, adapting them with facial emotion datasets to augment FER accuracy.TL techniques were employed to fine-tune the models with facial emotion data, resulting in heightened accuracies on datasets like JAFFE and KDEF. Challenges encountered in FER encompassed feature extraction, model construction, and data preprocessing, which were effectively addressed through deep learning methodologies and TL. The study underscored the advantages of TL, including reduced costs, hardware requirements, and increased accuracy in FER, emphasizing its significance in advancing emotion recognition technology [12].

The study delves into the realm of facial expression recognition by proposing a novel loss function tailored for Convolutional Neural Network (CNN) architectures. In pursuit of enhancing recognition accuracy, the researchers embarked on training various CNN architectures, including AlexNet, InceptionNet, and ResNet, utilizing different loss functions for evaluation purposes. Benchmark databases such as MMI, Oulu-CASIA, and FER2013 served as the testing grounds for assessing the efficacy of the proposed loss function against existing ones. Through meticulous experimentation and analysis, the proposed loss function demonstrated superior performance in terms of recognition accuracy across these benchmark datasets.

The paper begins with a comprehensive survey of previous loss functions employed in facial expression recognition tasks, providing valuable insights into the evolution of methodologies in this domain. Subsequently, the proposed loss function is elaborately described, highlighting its innovative features and potential advantages over existing approaches. Simulation results are meticulously analyzed, shedding light on the comparative performance of different CNN architectures trained with various loss functions. The conclusions drawn from these analyses serve to underscore the significance and effectiveness of the proposed loss function in advancing the state-of-the-art in facial expression recognition.

In summary, the paper presents a systematic exploration of loss functions in CNN architectures for facial expression recognition, culminating in the development and validation of a novel approach that surpasses existing methodologies in terms of recognition accuracy. Through rigorous experimentation and analysis, the study contributes valuable insights and methodologies to the field, paving the way for further advancements in facial expression recognition technology [13].

The study addresses Facial Expression Recognition (FER) challenges in real-time scenarios by leveraging ensemble methods with pre-trained deep learning architectures like AlexNet, ResNet50, and Inception V3. Emphasizing the significance of advanced deep-learning architectures, the research tackles FER using the FER2013 dataset, comprising grayscale images normalized to 48x48 pixels and associated with seven emotions. Given the dataset's small size and class imbalances, transfer learning is employed to enhance model accuracy, accompanied by preprocessing steps such as image resizing. The study contributes to the literature by showcasing the efficacy of ensemble methods and transfer learning in improving FER performance, particularly in real-time applications where accuracy and efficiency are paramount [14].

The study focuses on enhancing facial expression recognition accuracy in low-resolution images by integrating a voting mechanism into a residual network architecture. By processing multiple patches of each image through a single network, the method effectively determines the image's class while maintaining a constant number of trainable parameters. In the broader field of facial expression recognition, researchers have explored diverse approaches such as filter-based subspace learning, denoising techniques, super-resolution algorithms, and knowledge distillation. These methodologies aim to address challenges related to image quality, noise, and information extraction, ultimately contributing to advancements in accurate and robust facial expression recognition systems. The proposed approach adds to this body of literature by introducing a novel mechanism specifically tailored for improving accuracy in low-resolution image scenarios [15].

1.3 Problem Statement

The problem statement revolves around addressing the pressing need for considering temporal features alongside spatial features while prediction of emotions. Here's an expanded version of the problem statement. The existing system uses different architectures of Convolutional Neural Networks (CNN) for feature extraction and classification. The system primarily focuses on static images and mentions video data in a limited context. However, real-world scenarios often involve dynamic changes in facial expressions.

To address this challenge of neglecting temporal dependencies while predicting emotions, our research aims to develop a hybrid model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTMs). The primary objective is to consider temporal dependencies alongside special features for categorizing emotion of a person in a real time video. Our model seeks to overcome the limitations observed in the existing systems.

Temporal dependencies: Temporal dependencies refer to the relationships and connections between information across different points in time. In simpler terms, it's the "how things change over time" aspect of your data. Spatial features: Spatial features represent the characteristics of an object or region within a single frame. In video data, it's about the "what things look like at a specific point in time" aspect.

LSTMs are employed in our model after convolutions --which give a good knowledge of temporal dependencies for the model after capturing spatial features for every frame. Also, our hybrid model is capable of capturing several faces from a frame using the help of Haar-Cascade algorithm which makes our work easier to detect emotions of several people in a single frame.

Classification of Emotions: We have considered six major emotions Happy, Neutral, Sad, Fear, Angry, Surprise. After our model learns from the training data we use SoftMax logistic function at the output layer.

Addressing Limitations of Current Systems: Convolution Neural Networks have been employed to analyse facial expressions in the existing systems which are mainly of static based and they only can recognise emotions from static images. Different pre-trained models such as VGG16, DenseNet, EfficientNetB2, ResNet50, XceptionNet are used to train the model which gained accuracy in between 65% to 76%.

1.4 Objectives of the research

1.4.1 Capturing temporal dependencies:

Our primary objective is to capture temporal dependencies between frames to get a better insight of changes in the facial clues during a particular expression. These individual changes in facial clues, spread across different frames, are connected and tell a story of a particular emotion. Capturing these relationships is crucial for understanding the overall emotional state.

Emotion	Motion of Facial Parts
Happy	Open eyes, open mouth, lip corner pulled, cheeks raised
Sad	Outer eyebrow down, inner eyebrows raised, eyes closed, lip corner down
Surprise	Eyebrow up, open eyes, jaw dropped
Anger	Eyebrow pulled down, open eyes, lip tightened
Fear	Outer eyebrow down, inner eyebrow up, mouth open
Neutral	Face, lips, eyebrows at normal position

Table 1.4.1.1 Temporal Fetures

1.4.2 Identifying multiple faces in a single frame:

We have made use of Haar-Cascade algorithm to detect objects in images, irrespective of their scale in image and location. One of the primary benefits of Haar cascades is that they are just so fast. Haar-Cascade was employed to detect faces in our model. The haarcascade_frontal_face.default.xml file is used to detect faces.

1.4.3 Facial emotion recognition in a real-time video:

Facial emotion recognition has not only been limited for static image but also, we can use them for predicting emotion in a real-time video. This can be possible by the OpenCV algorithm where we can take feed directly from user's webcam. Our model now process video frame by frame capturing temporal dependencies among frames and spatial features in a frame and predicts emotions.

1.5 Databases Description

A dataset is a collection of data. Most commonly a dataset corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the dataset in question. The data set lists values for each of the variables such as the height or weight of an object for each member in the dataset. A data set is organized into some type of data structure. In a database, for example, a data set might contain a collection of business data (names, salaries, contact information, sales figures, and so forth). The database itself can be considered a data set, as can bodies of data within it related to a particular type of information, such as sales data for a particular corporate department.

Our dataset is a custom-made dataset taking images from CK+ (Cohn-Kanade) and FER2013 datasets was available in the Kaggle repository.

Our dataset contains grayscale images of faces displaying various facial expressions, such as:

1. Happy
2. Sad
3. Neutral
4. Fear
5. Angry

6. Surprise

We have considered these emotions because of the fact that these emotions are likely to appear on most of faces in the real-world scenario. Each image is labeled with the corresponding emotion. The below are the count of number of samples present in the dataset. We have made our model learn from these samples splitting the dataset into 70% and 30% for train and test respectively.

The below are the sample images of different emotion classes available in the dataset.

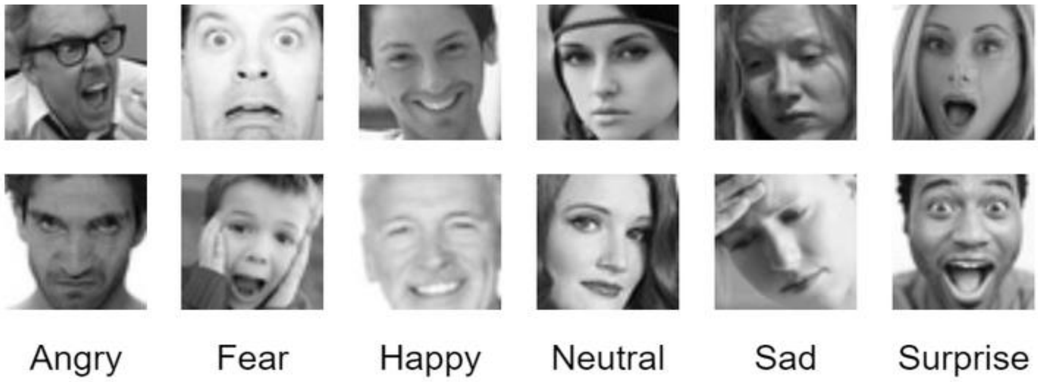


Fig 1.5.1 Dataset Samples

1.6 Similarity Methods Used

1.6.1 Categorical Cross Entropy

In this project, we have strategically incorporated the categorical cross entropy as a cornerstone similarity measure. This choice reflects our commitment to employing cutting-edge techniques in deep learning to enhance the performance and robustness of our model, particularly in the domain of multi-class classification tasks. Categorical cross entropy stands as a versatile and powerful tool in the arsenal of deep learning practitioners. Its utility lies in its ability to quantify the dissimilarity between the probability distributions predicted by our model and the true distribution encapsulated by the ground truth labels. By evaluating this disparity, categorical cross entropy serves as a compass guiding the optimization process during model training.

During the iterative training phase, our objective was to minimize the categorical cross entropy loss. This endeavor involved meticulous fine-tuning of the model's parameters to progressively align its predictions more closely with the true distribution of classes present in the training data. Through this iterative optimization, our model underwent a transformative journey, evolving to discern and capture intricate patterns inherent in the data, thereby enhancing its classification prowess. Beyond its role in model training, categorical cross entropy remains instrumental during the inference phase. Equipped with the insights gleaned from the training process, our model demonstrates its proficiency in making confident predictions on unseen data, effectively generalizing its learnings to real-world scenarios. This capability underscores the robustness and reliability of our approach, empowering our solution to thrive in diverse and dynamic environments. In essence, the integration of categorical cross entropy as a central similarity measure epitomizes our dedication to leveraging advanced methodologies to achieve superior performance in our project. By embracing state-of-the-art techniques in deep learning, we endeavor to push the boundaries of what is achievable, ultimately delivering impactful and innovative solutions that address real-world challenges with efficacy and precision.

1.6.2 Adam Optimizer

The Adam optimizer, a prevalent algorithm in deep learning, efficiently updates neural network weights during training by blending RMSprop and momentum techniques. It dynamically adjusts learning rates for each parameter based on gradient magnitudes and history, promoting faster convergence, and handling sparse gradients, especially beneficial for large datasets. By individually adapting learning rates, Adam optimizes training by adjusting step sizes for different gradient magnitudes, mitigating overshooting and local minima issues, resulting in swifter and more stable convergence. Its efficacy in optimizing intricate neural network architectures makes Adam widely preferred in deep learning, enhancing convergence speed and training efficiency.

1.7 Performance Evaluation Measures

In evaluating the efficiency of our hybrid deep learning model for dynamic emotion recognition, we employ a comprehensive set of performance metrics, including both training and validation metrics, to assess the model's effectiveness and generalization capabilities. These metrics provided insights into the model's ability to accurately classify images across the six different emotion categories: Happy, Sad, Neutral, Angry, Fear, Suprise.

1.7.1 Training Loss and Accuracy:

Training Loss: The training loss metric measures the discrepancy between the predicted outputs of the model and the ground truth labels during the training phase. A lower training loss indicates that the model is effectively learning to minimize errors and optimize its parameters to better fit the training data.

Training Accuracy: The training accuracy metric quantifies the percentage of correctly classified samples within the training dataset. It provides insight into how well the model is performing on the training data and serves as a proxy for the model's ability to learn and discriminate between different malware classes.

1.7.2 Validation Loss and Accuracy:

Validation Loss: The validation loss metric evaluates the performance of the model on a separate validation dataset that the model has not seen during training. It measures how well the model generalizes to unseen data and helps identify overfitting or underfitting issues. A lower validation loss indicates that the model is effectively capturing underlying patterns in the data without overfitting.

Validation Accuracy: The validation accuracy metric assesses the percentage of correctly classified samples within the validation dataset. It provides a measure of the model's ability to generalize to new, unseen data and serves as a crucial indicator of its real-world performance.

By monitoring both training and validation metrics, we can gauge the model's training progress, detect potential issues such as overfitting or underfitting, and make decisions regarding model architecture, hyperparameters, and optimization strategies. These performance metrics provide a comprehensive evaluation of the effectiveness and robustness of our hybrid deep learning model for emotion recognition.

CHAPTER-2

DYNAMIC EMOTION RECOGNITION USING HYBRID DEEP LEARNING MODELS

CHAPTER - 2

DYNAMIC EMOTION RECOGNITION USING HYBRID DEEP LEARNING MODELS

2.1 Brief Outline of the Chapter

Previous research in Facial Emotion Recognition has extensively explored the application of various machine learning techniques, including neural networks.

Neural networks, particularly convolutional neural networks (CNNs) and different pre-trained models such as VGG16, DenseNet, EfficientNetB2, ResNet50, XceptionNet, are used for feature extraction and classification.

These networks utilize layers of interconnected processing elements to extract meaningful features from input data (grayscale images). Additionally, convolutional layers, pooling layers, fully connected layers and LSTM layers are involved to play crucial roles in processing data, enabling the model to capture temporal dependencies and spatial features.

We propose a dynamic approach for facial emotion recognition using tour custom dataset. While pre-trained CNN models like VGG16 or InceptionV3 can be fine-tuned for emotion recognition, they have limitations. These models are often trained on large datasets of static images, and their features might not be optimal for capturing the dynamics of emotions in video. Our approach of training a CNN specifically for our task allows us to extract features that are more tailored to the problem of video-based emotion recognition.

Following the CNN stage, our model incorporates a Long Short-Term Memory (LSTM) network. LSTMs are specifically designed to handle sequential data like video frames. They possess a unique architecture that allows them to learn long-term dependencies within the data. In our case, the LSTM takes the sequence of features extracted by the CNN from individual frames and analyzes how these features evolve over time. This enables the model to capture the temporal dynamics of emotions and understand the relationships between changes in facial expressions.

The combination of CNN and LSTM in our hybrid model leverages the strengths of both architectures. The CNN extracts informative features from individual frames, while the LSTM captures the temporal relationships between these features.

This synergy allows the model to achieve superior performance in recognizing emotions from video data compared to relying solely on CNNs or LSTMs. The proposed model aims to improve the efficiency and accuracy of emotion recognition.

Also, multiple facial emotion detection can be possible because we make use of Haar-Cascade object detection algorithm to detect faces. This approach allows the model to recognize and classify emotions for multiple individuals present in the video stream simultaneously. This capability can be valuable in various applications, such as group video conferencing where the system recognizes emotions of all participants, or educational settings where emotions of multiple students can be detected during lectures. This paves the way for innovative applications that leverage the power of video-based emotion recognition.

2.2 Proposed Method

2.2.1 Neural Networks

A neural networks computing system made up of several simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. Neural networks are typically organized in layers. Layers are made up of several interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output.

2.2.2 Convolutional Neural Networks (CNN)

In traditional feed-forward neural networks, each neuron in the input layer is connected to every output neuron in the next layer – we call this a fully connected (FC) layer. However, on CNN, we don’t use FC layers until the very last layers in the network. We can thus define a CNN as a neural network that swaps in a specialized “convolutional” layer in place of a “fully-connected” layer for at least one of the layers in the network.

A nonlinear activation function, such as ReLU, is then applied to the output of these convolutions and the process of convolution and activation continues along with a mixture of other layer types to help reduce the width and height of the input volume and help reduce the width and height of the input volume and help reduce overfitting until we finally reach the end of the network and apply one or two FC layers where we can obtain our final output classifications.

Each layer in a CNN applies a different set of filters, typically hundreds or thousands of them, and combines the results, feeding the output into the next layer in the network. During training, a CNN automatically learns the values for these filters. In the context of image classification, our CNN may learn to:

- Detect edges from raw pixel data in the first layer.
- Use these edges to detect shapes in the second layer.
- Use these shapes to detect higher-level features in the highest layers of the network.

The last layer on CNN uses these higher-level features to make predictions regarding the contents of the image. In practice, CNNs give us two key benefits: local invariance and compositionality. The concept of local invariance allows us to classify an image as containing a particular object regardless of where in the image the object appears. We obtain this local invariance through the usage of “pooling layers” which identifies regions of our input volume with a high response to a particular filter.

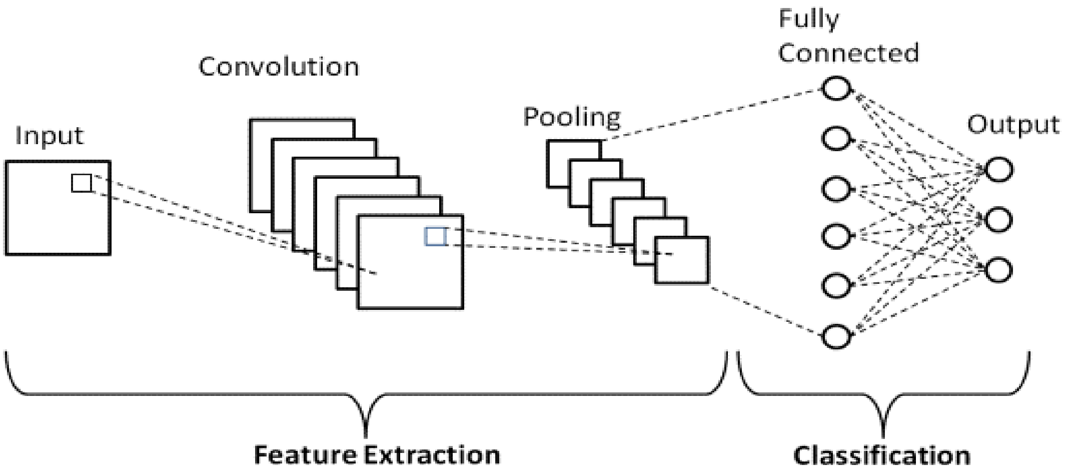


Fig 2.2.2.1 CNN architecture

Convolutional Neural Networks (CNNs) have shown promising results in emotion recognition. But, CNNs typically process video frames independently, neglecting the temporal relationships between frames. They miss the crucial information about how an expression evolves over time, which is essential for accurate emotion recognition.

Following the CNN layer, LSTMs process the sequence of feature vectors extracted by the CNN. They analyze the temporal relationships between these vectors, capturing the dynamics of facial expressions over time. This allows the model to understand how an emotion unfolds and intensifies/weakens within the video.

Input representation: The input images are converted into pixels range from 0 to 255. Each image is of (48,48,1) shape with only a single channel (grayscale images). Grayscale images are used for the training of the model because they have only one channel as opposed to three channels which reduces the size of each image and the overall data set leading to faster processing times and lower storage requirements. Also, sometimes colour can be distracting from the underlying shape and texture of facial features which are crucial for emotion recognition, grayscale images eliminate this distraction.

Pre-processing: The input for training the model is an image. They have to be converted to a 3-dimensional image of size (48*48*1). We will normalize the image arrays this is done because neural networks are highly sensitive to non-normalized data. We will use min-max normalization.
For these gray-scaled images min=0, max=255 therefore we will divide the array by 255 because,

$$x_{min-max} = \frac{x - min(X)}{max(X) - min(X)}$$

Convolutional layers: The convolutional layers of the CNN learn spatial patterns in the input. The filters in the convolutional layers slide over the input and learn features such as edges, corners, and blobs. We have applied [padding = same] to make sure the image size doesn’t reduce further when it goes through convolutions. Detect edges from raw pixel data in the first layer and shapes in the second layer and shapes in the highest layers of the network.

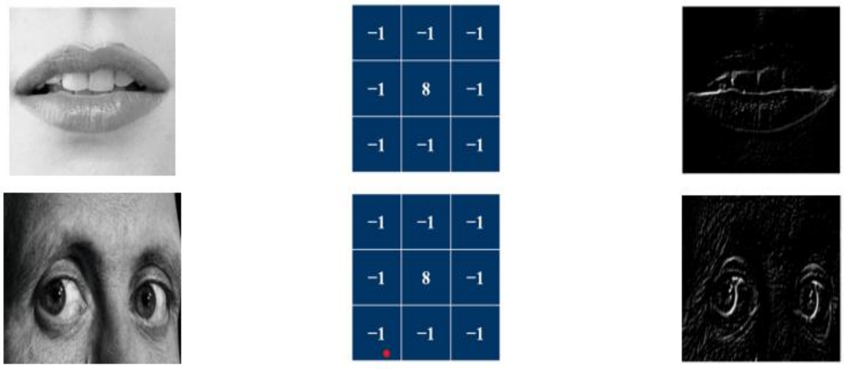


Fig 2.2.2.2 Kernel

Pooling layers: In Convolutional Neural Networks (CNNs), pooling layers play a crucial role alongside convolutional layers in the feature extraction process. Pooling layers down sample the feature maps generated by convolutional layers. This reduces the spatial dimensions (height and width) of the data, leading to several benefits:

Reduced Model Complexity**:** Fewer parameters are needed to represent the data, making the model more efficient and easier to train, especially for large datasets.

Less Overfitting: By reducing the amount of data, pooling layers can help to prevent the model from overfitting to the training data and improve its ability to generalize to unseen examples. The pool size determines the level of down sampling. Larger pool sizes reduce the dimensionality more significantly but might also lose more spatial information. We have used Max pooling in our model because identifying the most dominant features is important.

Dropout layers: The dropout layers of the CNN are a type regularization technique used to prevent overfitting by randomly deactivating a fraction of neurons during training.

Flatten layers: The flatten layer's job is to transform the data from a 2D format into a 1D format, which is necessary before giving them to the LSTM units.

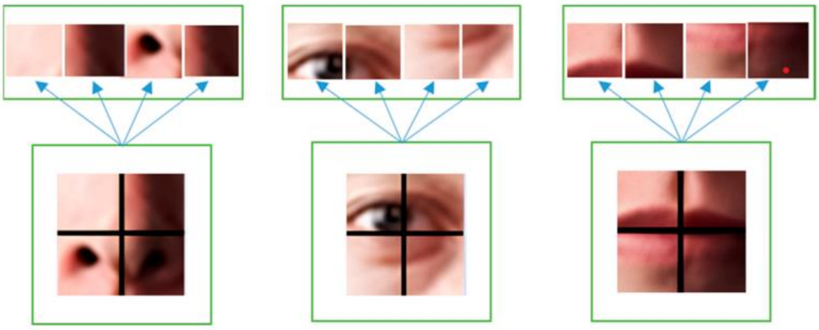


Fig 2.2.2.3 Flattening a feature vector

2.2.3 Long-Short-Term-Memory (LSTM)

Long-Short-Term-Memory (LSTM) unit is made up of four feedforward neural networks. Three of the four feedforward neural networks are responsible for selecting information. They are the forget gate, the input gate, and the output gate. These three gates are used to perform the three typical memory management operations: the deletion of information from memory (the forget gate), the insertion of new information in memory (the input gate), and the use of information present in memory (the output gate). The fourth neural network, the candidate memory, is used to create new candidate information to be inserted into the memory.

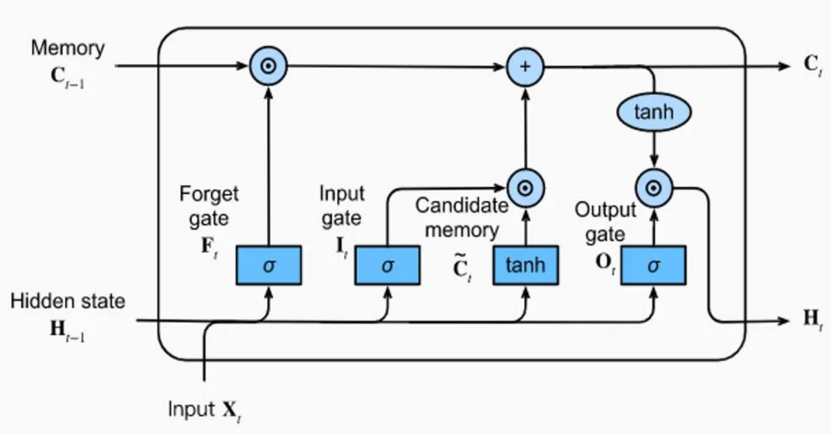


Fig 2.2.3.1 LSTM architecture

Fully connected layers: The output from the LSTM layer is flattened and fed to fully connected layers, which learn the mapping between the learned features and the emotion classes. In Convolutional Neural Networks (CNNs), fully connected (FC) layers, also sometimes called dense layers, come into play after the convolutional and pooling layers. They mark a shift from processing spatial information in feature maps to high-level reasoning and classification. Convolutional and pooling layers excel at extracting spatial features from the input data (images or video frames). They identify edges, lines, shapes, and other visual elements within localized regions. Fully connected layers, on the other hand, process these extracted features in a more global manner.

Output: The SoftMax function takes a vector of real numbers as input (typically the output scores from the last fully connected layer in a CNN). These scores represent the network's "confidence" in each potential class for the input data (e.g., different emotions in your video emotion recognition model). The SoftMax function then applies a mathematical transformation to these scores, converting them into a probability distribution between 0 and 1. The sum of all these probabilities will always be 1. The final output of the model is the emotion with highest probability distribution over the emotion classes.

2.2.4 Haar-Cascade object detection algorithm

Haar Cascade is particularly well-suited for facial detection because the Haar-like features can be used to distinguish between facial features such as the eyes, nose, and mouth. The algorithm detects faces by first creating a Haar Cascade classifier using a set of positive and negative images. The positive images contain faces, while the negative images do not. The classifier is then used to scan a new image or video for faces. The scanning process involves sliding a window of fixed size over the image and applying the classifier to each window. If the classifier detects a face in the window, it is marked as a potential face. The potential faces are then filtered based on their size, position, and shape to reduce false positives.

To implement Haar Cascade in Python, we will use the OpenCV library, which provides pre-trained Haar Cascade classifiers for facial detection. We will use the cv2.CascadeClassifier class to create a classifier, and the detectMultiScale method to detect faces in an image.

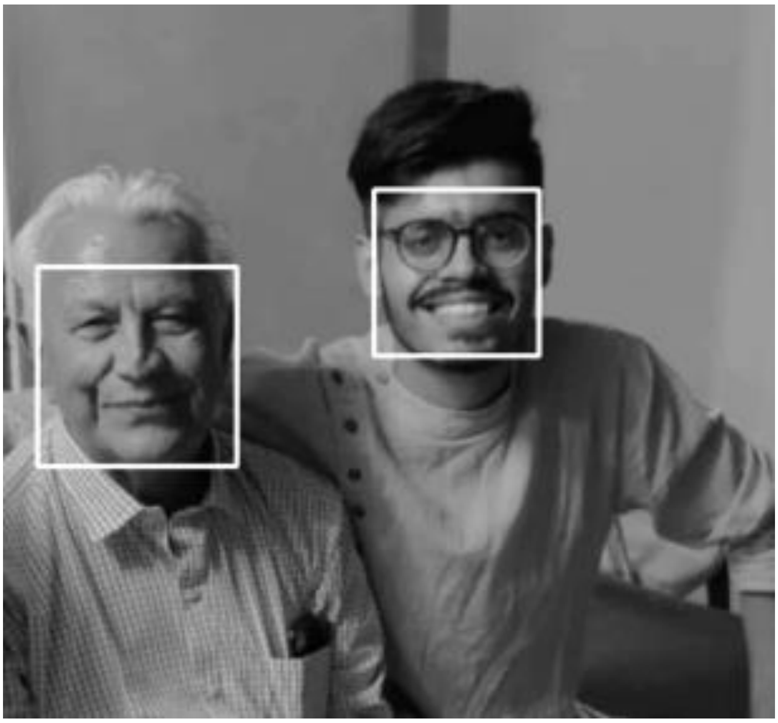


Fig 2.2.4.1 Faces identified by Haar Cascade

2.2.5 Proposed Hybrid CNN-LSTM architecture

The proposed model consists of grayscale images with a single channel, these images are preprocessed and deep learning features are implemented for the classification of emotions. The methodology is a hierarchy starting from the data collection. Here the data is collected from the open source Kaggle website and then made into a custom dataset, then the collected samples are preprocessed followed by CNN-LSTM hybrid architecture that is used for spatial feature extraction taking care of temporal dependencies as well as classification of emotions using SoftMax layer. Finally, the results are analyzed using performance metrics such as Accuracy

Our dataset consists of grayscale images with a single channel. Six classes of images are present in the data, both in the training as well as testing set. It is embossed with six classes of emotions such as Happy, Sad, Angry, Fear, Neutral, Suprise.

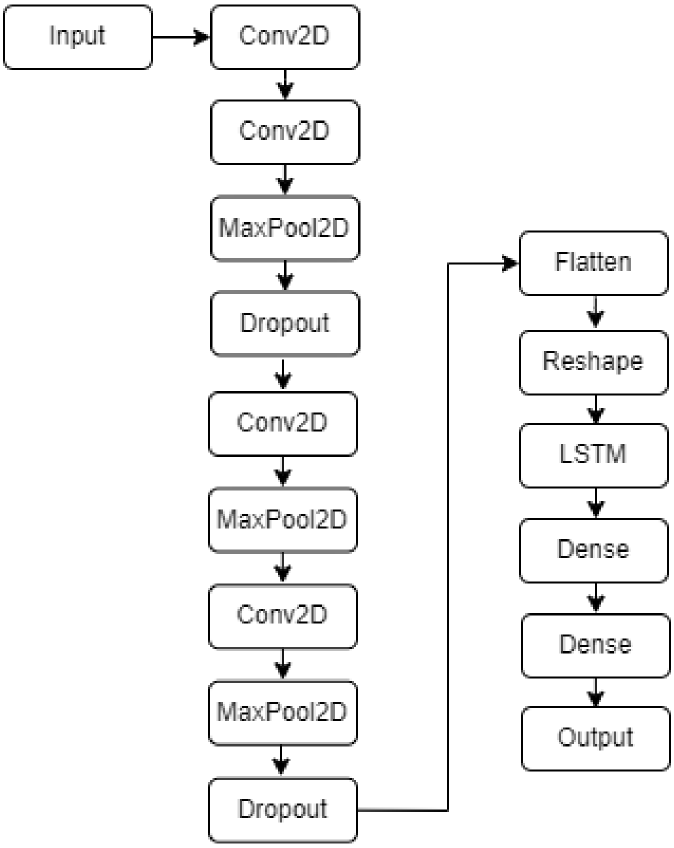


Fig 2.2.5.1 Model Architecture

2.2.6 CNN-LSTM Algorithm steps for Dynamic Facial Emotion Recognition

```
Step 1:  Start webcam access
Step 2: Consider a frame (i)
Step 3: Plot boundary boxes of face through Haar-Cascade algorithm
Step 4: Preprocess the boundary box portion of frame.
Step 5: Input the boundary box portion to the hybrid (CNN-LSTM) model.
Step 6: Emotion prediction and display emotion.
Step 7: if user enter ‘q’: Stop
        else: Take next frame (i+1) and go to step 3
```

For the dynamic facial emotion recognition, we make use of OpenCV to capture feed from the webcam. The feed was split in frames and each frame was passed through the Haar-Cascade object detector to detect multiple faces. Those faces are sent to our hybrid model to predict emotions and was shown alongside face.

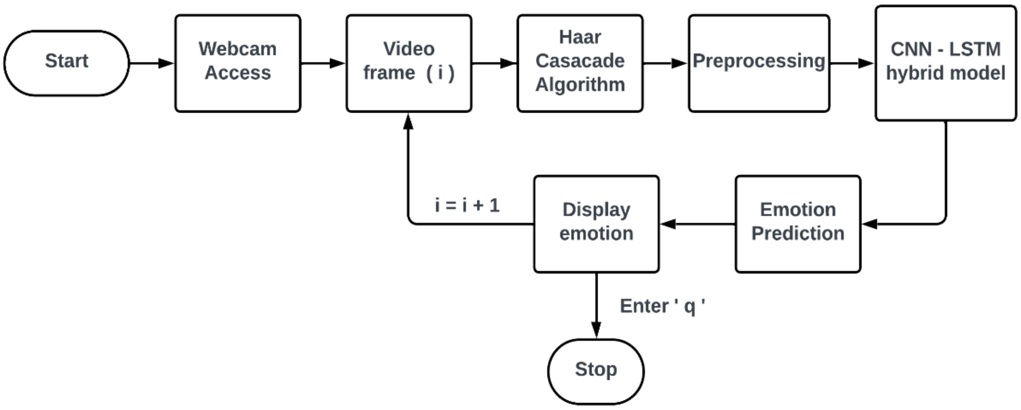


Fig 2.2.6.1 Work flow of our dynamic model

2.2.7 Sample Code

The below code was used to define the hybrid model for emotion recognition task.

model = Sequential()

# This line initializes a sequential model, which is a way to define neural networks in Keras

model.add(Conv2D(32, kernel_size = (3,3), activation='relu',input_shape=(48,48,1)))

# This layer applies a 2D convolution operation with 32 filters, each of size 3x3, to the input image. The input shape is specified as (48, 48, 1). The rectified linear unit (ReLU) activation function is applied to the output of the convolution, introducing non-linearity.

model.add(Conv2D(64, kernel_size = (3,3), activation='relu'))

model.add(MaxPooling2D(2,2))

# This layer performs max pooling with a pool size of 2x2. It down samples the feature maps by taking the maximum value within each 2x2 region, reducing the spatial resolution but potentially strengthening the extracted features.

model.add(Dropout(0.25))

# Dropout layers randomly drop a certain percentage (25% in this case) of units during training. This helps prevent overfitting.

model.add(Conv2D(128, kernel_size = (3,3), activation='relu'))

model.add(MaxPooling2D(2,2))

model.add(Conv2D(128, kernel_size = (3,3), activation='relu'))

model.add(MaxPooling2D(2,2))

model.add(Dropout(0.25))

```
model.add(Flatten())
# The model here converts the extracted feature maps into a one-dimensional vector suitable for feeding into the LSTM layer.
model.add(Reshape((1,-1)))

model.add(LSTM(64))
# This line of code introduces an LSTM layer with 64 memory units. The LSTM layer is responsible for capturing temporal dependencies within the flattened feature vector.
model.add(Dense(1028,activation='relu'))
model.add(Dense(6,activation='softmax'))
# The final layer has 6 neurons (likely corresponding to the 6 emotions your model classifies) and uses the softmax activation function. Softmax ensures the output probabilities sum to 1, making it suitable for multi-class classification.
```

The below code was used to define random transformations to the data which in terms lead to the model encounter wider variety of image variations. This helps it learn features that are more generalizable and less prone to overfitting on the specific training set. For example, if your training images only contain faces looking straight ahead, the model might struggle to recognize faces turned slightly to the side. By introducing random rotations, the model learns to recognize features regardless of the head orientation.

```
train_datagen_1 = ImageDataGenerator(
rotation_range=10,
# This argument introduces random rotations between -10 and 10 degrees
width_shift_range=0.1,
# This argument applies random horizontal shifts of up to 10% of the image width.
height_shift_range=0.1,
# This argument applies random vertical shifts of up to 10% of the image height.
shear_range=0.1,
# This argument introduces random shearing transformations of up to 10% in either direction.
zoom_range=0.1,
# This argument applies random zoom of up to 10% in or out of the image.
horizontal_flip=True,
# This argument enables random horizontal flipping of the images, creating variations where left and right sides are reversed.
)
```

The below code was used to train our defined CNN-LSTM model taking training and validation data.

```
history = model.fit(
train_datagen_1.flow(X_train, y_train, batch_size =20),
# The flow method here provides a generator that continuously iterates over our training data while applying the defined data augmentations.
validation_data=(X_test, y_test),
epochs=75,
)
```

Code to Make Model Dynamic

```
emotion_labels=['Angry','Fear','Happy','Neutral','Sad','Surprise']
# Defining labels for emotions
cap=cv2.VideoCapture(0)
# Initializing a video capture object using the default webcam (index 0).
while True:
   if not cap.isOpened():
     print("Error: Video capture object is not opened.")
     exit()
    _, frame=cap.read()
# Reading the frames captured by the webcam one by one through loop.

faces=face_classifier.detectMultiScale(gray_frame)
# Detection of faces in the frame mentioned using the haar-cascade object and stores the co-ordinates of each face present in the faces variable.

prediction=classifier.predict(roi)[0]
```

Emotion prediction of faces available in the region of interest (roi) with the help of our saved model(classifier).

2.3 Results and Discussions:

The results of our study showcase the potential of utilizing convolutional neural network (CNN) and Long Short-Term Memory (LSTM) hybrid architecture for the recognition of emotions considering spatial as well as temporal dependencies. The high accuracy was achieved by our proposed system rather than the pre-trained convolutional network models. Our model has acquired ~92% of accuracy over 75 epochs which is a better sign. By leveraging webcam access using OpenCV it was made dynamic and easy to use for any person. It will make our dynamic nature of model robust in such a way that it can predict the emotions of persons in real-time with minimum error-rate. Furthermore, the robust performance of our model in predicting emotions highlights its potential in several areas. Moving forward, continued research efforts are warranted to further refine and validate our proposed system, including the exploration of larger and more diverse datasets, the integration of multiple modalities such as speech and text.

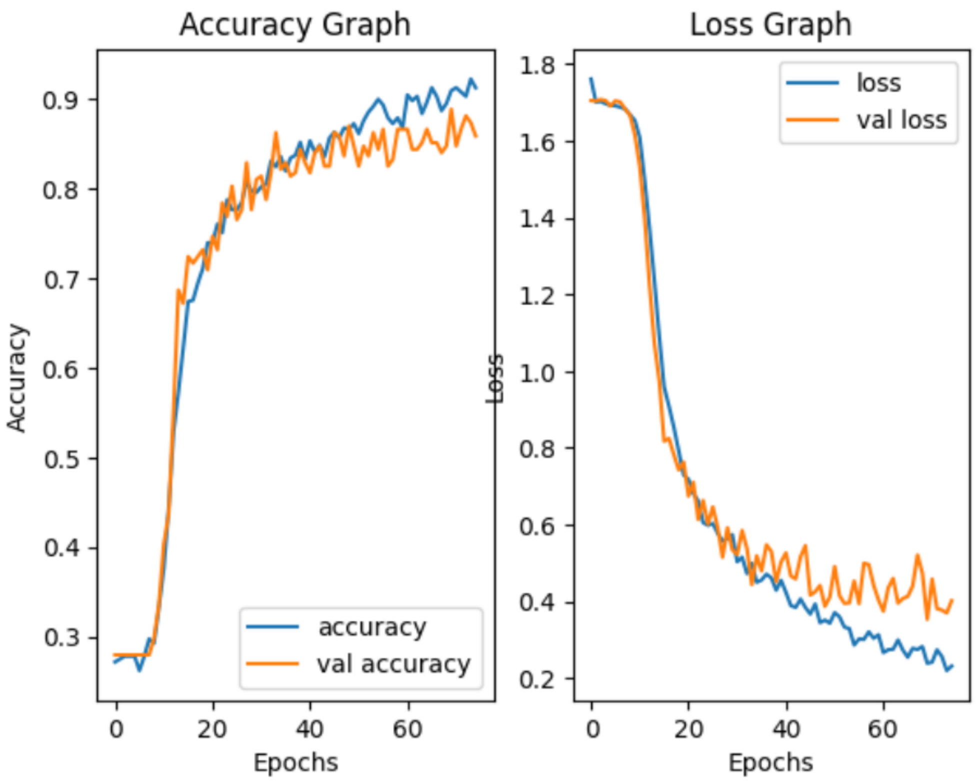


Fig 2.3.1 Accuracy and Loss graphs

2.3.1 Comparison with other models

Pre-trained models like VGG16, InceptionV3, or ResNet architectures can be fine-tuned for emotion recognition. These models excel at extracting informative features from individual video frames. Also, CNN based architectures excel at extracting spatial information. The accuracy was adopted as the evaluation metrics. Both our prosed and some of our existing models have been trained using the Facial Emotion Recognition dataset available in the Kaggle repository.

Our hybrid model outperforms the pre-trained models and CNN based architectural models with accuracy around ~90%. The below table provides a detailed view of accuracy comparison between our proposed model and pre-trained models mentioned in our reference papers.

Models	**Accuracy (%)**
5-layer CNN	67.69
6-layer CNN	67.63
RESNET50	72.72
VGG16	70.22
Xception	65.19
DCNN	66.75
EfficientNetB2	68.09
DenseNet	69.10
Proposed Method (CNN-LSTM)	92.07

Table 2.3.1.1 Accuracy comparison

From table 2.3.1.1, we can conclude that the proposed methodology yields comprehensive experimental results that demonstrate the superiority of the hybrid CNN-LSTM approach over the pre-trained models.

Test Cases

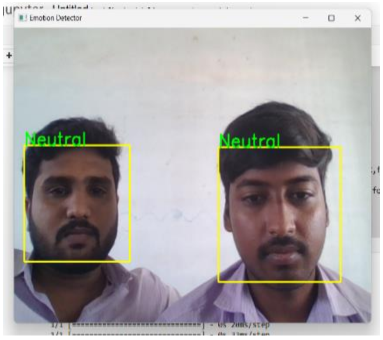


Fig 2.3.1.1 Neutral Testing Image

Fig 2.3.1.1 has person with normal lips, eyebrows and mouth which made model to predict as “Neutral”.

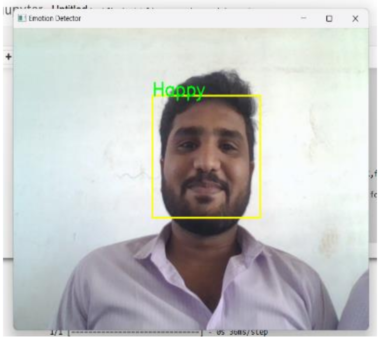


Fig 2.3.1.2 Happy Testing Image

Fig 2.3.1.2 has person with eyes open, lip corners pulled and cheeks raised which made model to predict as “Happy”.

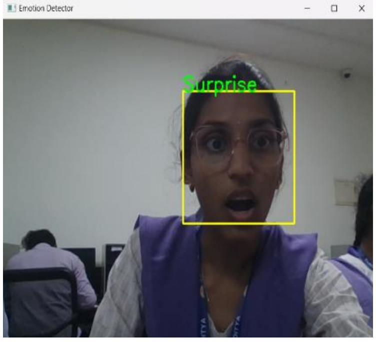


Fig 2.3.1.3 Surprise Testing Image

Fig 2.3.1.3 has person with jaw dropped, open eyes and eyebrows up which made model to predict as “Suprise”

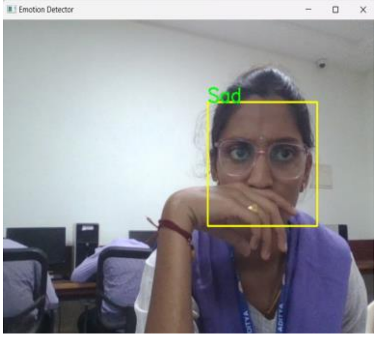


Fig 2.3.1.4 Sad Testing Image

Fig 2.3.1.4 has person with inner eyebrows down which made model to predict as “Sad”.

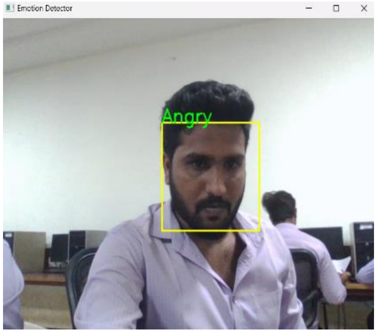


Fig 2.3.1.5 Angry Testing Image

Fig 2.3.1.5 has person with lips tightened, eyes open and eyebrow pulled down which made model to predict as “Angry”.

2.4 The Main Contribution of the Chapter

The main contribution of this project is that our model can capture the changes in the facial clues between frames and recognize emotion based on those changes which make model robust and more accurate. This comprehensive examination not only highlights the its usage in emotion recognition but also its application of capturing a particular sequence in a video. To begin with, the chapter provides a meticulous breakdown of CNN architecture, elucidating the intricacies of convolutional layers, pooling layers, and fully connected layers. By delving into these components, readers gain a profound comprehension of how CNNs operate on facial images to find spatial features and LSTMs on temporal dependencies.

Moreover, the discussion extends to the practical aspects of utilizing CNNs and LSTMs in a finding something over a sequence of frames in a video. The chapter underscores the importance of curated datasets with diverse samples to train our model effectively, ensuring robust model performance across different populations and imaging variations. In addition to model architecture and dataset considerations, the chapter dives into temporal based feature extraction techniques tailored a sequence of images. It explores methods for capturing relevant insights called as temporal dynamics among a sequence of frames in a video.

Furthermore, the chapter makes use of dynamic nature of recognizing emotions in a real time video. It outlines approaches for using OpenCV to access users webcam to access feed. Dynamic access of this model provides help in various fields of study. Detecting emotions for multiple faces at a single instant or a frame makes the model more robust and effective.

In summary, the chapter's main contribution lies in bridging temporal dependencies and spatial features for dynamic emotion recognition of multiple faces.

2.5 Conclusions

In this project, a hybrid CNN-LSTM deep learning model was developed and trained to predict the emotions. The model was trained and tested on a custom dataset made from facial expression datasets available in the Kaggle repository.

The model has identified emotions with good accuracy when faces are accessed dynamically through webcams of user. Multiple faces are also identified and emotions of every face was identified. Our hybrid model works very well for static images as well as dynamic video. Our model has achieved training accuracy around ~92% and testing accuracy around ~86% for 75 epochs.

This is a better outcome as it would lead to better understanding of emotions or a particular action over a sequence of time in a real-time scenario. Furthermore, taking this model to the next level makes the use of this hybrid model not only for emotion recognition but also for analyzing a series of images (that we name “frames”) that are taken in time to capture a particular action.

CHAPTER-3

CONCLUSION AND FUTURE SCOPE

CHAPTER -3

CONCLUSION AND FUTURE SCOPE

3.1 CONCLUSION

This project explored the development of a hybrid CNN-LSTM model for video-based emotion recognition. We addressed the limitations of CNNs in capturing the temporal dynamics of emotions within video sequences. The proposed model leverages the strengths of both architectures: CNNs: Effectively extract informative spatial features (facial landmarks, wrinkles) from each video frame. LSTMs: Analyze the sequence of feature vectors, capturing the evolution of expressions over time.

This combined approach leads to a more comprehensive understanding of the emotional content in videos. Our findings demonstrate that the hybrid CNN-LSTM model achieves significantly higher accuracy in recognizing emotions compared to using CNNs alone.

Furthermore, we discussed the challenges associated with real-world webcam-based emotion recognition, including factors like lighting variations, user characteristics, and webcam limitations. We explored strategies to mitigate these challenges, such as data augmentation and user education. Additionally, we emphasized the importance of security measures like data anonymization, model explainability, and user control to ensure responsible development and deployment of this technology.

In conclusion, this project demonstrates the potential of hybrid CNN-LSTM models for accurate and robust video-based emotion recognition. By addressing the real-world challenges and prioritizing user privacy, this technology can have valuable applications in various domains, such as human-computer interaction, healthcare, and education.

3.2 FUTURE SCOPE

This project lays a strong foundation for future advancements in video-based emotion recognition using hybrid CNN-LSTM models. Here are some exciting possibilities to explore:

- Multi-modal Integration: Extend the model beyond visual cues by incorporating additional modalities like voice analysis and body language recognition. This can provide a richer understanding of emotions and lead to even more accurate predictions.
- Advanced Architectures: Investigate more sophisticated deep learning architectures like transformers or ensemble methods to potentially improve the model's performance and generalization capabilities.
- Personalization and Adaptation: Develop mechanisms for the model to adapt to individual users over time. This could involve learning personalized expression patterns and improving accuracy for specific individuals.
- Real-time Processing: Optimize the model for real-time processing, enabling applications like real-time sentiment analysis during video conferencing or emotional response measurement in educational settings.
- We also have scope of building a time-distributed model which takes a bunch of samples as input at a time and gives outputs. It will be more robust for data that are chronologically ordered.

Also, the dataset Facial Emotion Recognition (FER2013) has samples with label mismatch which tells a fact that the emotions can vary even with looks (i.e.) there could be a person being sad over fear or a person being happy over surprise. There could be a way to develop a model which identifies emotions with probabilities. By exploring these future directions, we can unlock the full potential of emotion recognition models, paving the way for more nuanced human-computer interaction, personalized experiences, and impactful applications in various fields.

BIBLIOGRAPHY

[1] Swarna Kuchibhotla, Aruna S, Hima Deepthi Vankayalapati, Analysis of Facial Emotion Recognition for Image and Video Data using Convolution Neural Networks

[2] Christian Białek 1, Andrzej Matiola ´nski 1,2, and Michał Grega, An Efficient Approach to Face Emotion Recognition with Convolutional Neural Networks

[3] Rit Lawpanom, Wararat Songpan and Jakkrit Kaewyotha, Advancing Facial Expression Recognition in Online Learning Education Using a Homogeneous Ensemble Convolutional Neural Network Approach

[4] Jui-Feng Yeh *, Kuei-Mei Lin, Chia-Chen Chang and Ting-Hao Wang, Expression Recognition of Multiple Faces Using a Convolution Neural Network Combining the Haar Cascade Classifier

[5] Muhammad Abdullah, Mobeen Ahmad, Dongil Han, Facial Expression Recognition in Videos

[6] Apeksha Khopkar1 and Ashish Adholiya Saxena2, Facial Expression Recognition Using CNN with Keras

[7] Himaja Avula, Ranjith R, Dr. Anju S Pillai, CNN based Recognition of Emotion and Speech from Gestures and Facial Expressions

[8] Wassan Hayale, Pooran Singh Negi, and Mohammad H. Mahoor, Deep Siamese Neural Networks for Facial Expression Recognition in the Wild

[9] Mustafa Can Gursesli, Sara Lombardi, Mirko Duradoni, Leonardo Bocchi, Andrea Guazzini, And Antonio Lanata, Facial Emotion Recognition (FER) Through Custom Lightweight CNN Model Performance Evaluation in Public Datasets

[10] Renuka S. Deshmukh, Vandana Jagtap, Shilpa Paygude, Facial Emotion Recognition System through Machine Learning approach

[11] Dr. Shalini Gupta and Dr. Shubha Jain, Feeling Recognition by Facial Expression Using Deep Learning

[12] Usha Rawat, C.S. Rai, Improving Facial Emotion Recognition Through Transfer Learning with Deep Convolutional Neural Network (DCNN) Models

[13] Trong-Dong Pham, Minh-Thien Duong, Quoc-Thien Ho, Seongsoo Lee and Min-Cheol Hong, CNN-Based Facial Expression Recognition with Simultaneous Consideration of Inter-Class and Intra-Class Variations

[14] Resmi K. Reghunathan, Vineetha K. Ramankutty, Amrutha Kallingal and Vishnu Vinod, Facial Expression Recognition Using Pre-Trained Architectures

[15] José L. Gómez-Sirvent, Francisco López de la Rosa, María T. López and Antonio Fernández-Caballero, Facial Expression Recognition in the Wild for Low-Resolution Images Using Voting Residual Network