EMOTION THAT SPEAKS: PEERING BEYOND THE OBVIOUS WITH DEEP LEARNING FOR EMOTION RECOGNITION

BY

MD SHAKIL HOSSEN
ID: 203-15-14472

AND

NAZMUL ISLAM SHIMUL
ID: 203-15-14486

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Md. Sazzadur Ahamed
Assistant Professor
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

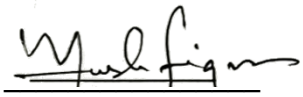
July 2024

APPROVAL

This Project titled "**Emotion that Speaks: Peering Beyond the Obvious with Deep Learning for Emotion Recognition**", submitted by **Md Shakil Hossen** and **Nazmul Islam Shimul** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on **15-07-2024**.

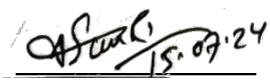
BOARD OF EXAMINERS

Narayan Ranjan Chakraborty — **Board Chairman**
Associate Professor & Associate Head
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Mushfiqur Rahman — **Internal Examiner 1**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Israt Jahan — **Internal Examiner 2**
Senior Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Dr. Abu Sayed Md. Mostafizur Rahaman — **External Examiner**
Professor
Department of Computer Science & Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Abdus Sattar, Assistant Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

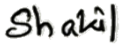
Supervised by:

Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Md. Sazzadur Ahamed
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:



Md Shakil Hossen
ID: 203-15-14472
Department of CSE
Daffodil International University

Nazmul Islam Shimul
ID: 203-15-14486
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Abdus Sattar, Assistant professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Human Computer Interaction*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Human emotions are spontaneous mental states produced by changes in facial muscles, leading to expressions. In various human-computer interaction applications, techniques for nonverbal communication like facial expressions, eye movements, and gestures are employed. Facial emotion, in particular, is widely utilized for conveying an individual's emotional states and feelings. However, emotion recognition is challenging due to the need for a clear distinction between facial expressions and the complexity and variability of emotions. Conventional machine learning algorithms frequently have difficulties in accurately recognizing emotions since they heavily depend on human-generated elements. To address this issue, we explored the use of deep learning models for emotion detection based on facial expressions. Specifically, we evaluated Vision Transformer (ViT), VGG19, InceptionV3, EfficientNet, and ResNet50 models. The findings of our study demonstrated that Vision Transformer (ViT) achieved the highest accuracy rate of 82.96%, followed by Efficient-Net at 82.36%, ResNet50 at 80.87%, InceptionV3 at 79%, and VGG19 at 78.22%. Based on its excellent accuracy and robustness, we propose using the Vision Transformer (ViT) for the identification of six distinct emotions: anger, neutrality, happiness, sadness, disgust, and surprise.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

FER	Facial Emotion Recognition
CNN	Convolutional Neural Network
ViT	Vision Transformer
ResNet50	Residual Network with 50 layers
VGG19	Visual Geometry Group Network with 19 layers
InceptionV3	Inception Network Version 3
EfficientNet	Efficient Convolutional Network
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
FERC	Facial Emotion Recognition Classifier
ANN	Artificial Neural Network
SHORE	Sophisticated High-speed Object Recognition Engine
AlexNet	Alex Network

CHAPTER 1

Introduction

1.1 Overview

Expressions on the face are essential for understanding and recognizing emotions. They have a significant influence on the dynamics of communication between parties, especially when it comes to HCI (human-computer interaction). The term "interface" takes on heightened significance as it underscores the crucial role the face plays in facilitating meaningful exchanges. Research in this domain suggests that the interpretation of facial expressions can wield a substantial influence on how spoken words are understood, essentially acting as a silent yet powerful communicator that shapes the nuances of a conversation. This underscores the centrality of emotional expression in interpersonal interactions, emphasizing its pivotal role in conveying nuanced messages beyond the mere spoken word. In the context of ideal Human-Computer Interfaces (HCI), there exists a genuine aspiration for machines not only to comprehend but accurately perceive and respond to human emotions.

This study delves into the intricate realm of emotion detection through the lens of a deep learning model, a sophisticated system designed to adeptly analyze facial images as a medium for detecting human emotion. The origins of this model can be traced back to Darwin's pioneering work on emotions, a foundation that has since captivated and galvanized numerous researchers to explore the intricate interplay between emotional states and facial expressions. The primary objective of this research is to present a robust and effective method for the detection of six fundamental emotions: anger, neutrality, happiness, sadness, surprise, and disgust, all rooted in the intricate tapestry of facial expressions. The field of emotion recognition using facial expressions has advanced significantly with deep learning models, particularly Vision Transformers (ViT) and transfer learning models. Vision Transformers bring a novel approach to image recognition by using self-attention mechanisms to process image patches. This allows ViTs to capture both local and global dependencies more effectively than traditional convolutional neural networks (CNNs), resulting in state-of-the-art performance in various tasks, including emotion recognition. On the contrast, Transfer learning enables models pre-trained on large datasets like ImageNet to be fine-tuned for specific tasks with smaller datasets. Models such as ResNet50, VGG19, InceptionV3, and Efficient-Net have shown great success in emotion recognition by leveraging pre-learned features, reducing the need for extensive computational resources and time.

1.2 Background and Present State

Facial expressions are crucial in human communication since they provide vital clues that improve the comprehension and interpretation of spoken words. The non-verbal aspect of communication is particularly vital in Human-Computer Interaction (HCI), as effective interfaces need to accurately convey the various intricacies of human emotions to enable meaningful interactions. Studies suggest that facial expressions substantially impact how communications are interpreted, serving as a potent nonverbal communicator that carries emotional meaning. Understanding these expressions is crucial for successful interpersonal communication, since they provide a significant amount of the signals that are dependent upon throughout talks. This field has led to significant progress in comprehending the intricate connection between emotional states and facial expressions. Within the field of Human-Computer Interaction (HCI), there is a significant aspiration for machines that can precisely recognize and react to human emotions. This would greatly enhance their ability to engage with humans more naturally.

The field of emotion detection has witnessed remarkable progress with the advent of deep-learning models designed to analyze facial images. These advanced systems have the objective of precisely detecting six fundamental emotions: anger, neutrality, pleasure, sorrow, surprise, and disgust. Among the cutting-edge techniques, Vision Transformers (ViT) and transfer learning models stand out for their efficacy. Vision Transformers employ self-attention processes to analyze picture patches, effectively capturing both local and global relationships more efficiently than conventional convolutional neural networks (CNNs). This novel methodology has resulted in exceptional achievement in several picture recognition assignments, including the identification of emotions. Meanwhile, transfer learning takes advantage of models pre-trained on large data sets such as ImageNet, which allows them to be fine-tuned for specific applications using smaller datasets.

1.3 Problem Statement

The study addresses the challenge of machines accurately understanding human emotions through facial expressions, crucial for advancing Human-Computer Interaction (HCI). Currently, the lack of a precise emotion detection method hinders seamless and emotionally intelligent interfaces, limiting meaningful human-computer communication. This shortfall hinders the development of seamless, emotionally intelligent interfaces, thereby restricting the depth and quality of human-computer communication. The primary problem lies in the need for sophisticated models, such as those powered by

deep learning, that can effectively discern and respond to the wide range of human emotions conveyed through facial cues. To close the gap and develop HCI solutions that are more meaningful, responsive, and intuitive, this issue must be resolved.

1.4 Objective

The inspiration for this research stems from a significant deficiency in contemporary technology, particularly in the field of Human-Computer Interaction (HCI), wherein robots frequently have difficulties in comprehending and reacting to human emotions as expressed through facial expressions. Emphasizing the crucial role of facial cues in communication, the research aims to contribute to the development of emotionally intelligent interfaces. The primary objective is to select an advanced emotion detection from deep learning models, which utilizes facial images to accurately recognize and categorize six fundamental emotions: anger, neutrality, fear, happiness, sadness, and surprise. Drawing insights from evolutionary theories, particularly Darwin's groundbreaking work, the study seeks to refine the conceptual framework of the proposed model. The overarching goal is to advance HCI by establishing a robust methodological framework for machines to interpret and respond to facial expressions, ultimately enhancing the overall quality of human-computer interactions. The research includes comprehensive evaluations to assess the accuracy and dependability of the proposed model across various scenarios and datasets. This endeavor not only addresses current limitations but also aims to direct upcoming studies in the dynamic field of emotion detection.

1.5 Scope and Limitations

The research intends to construct a strong deep-learning model capable of recognizing and categorizing six fundamental emotions such as anger, neutrality, disgust, happy, sad, and surprise using face images. It involves assessments across varied scenarios and datasets to assure practical relevance and accuracy in capturing complex facial emotions in actual human interactions. A fundamental aim is to increase Human-Computer Interaction (HCI) by building an effective framework for machines to comprehend and respond to facial emotions, thereby generating emotionally intelligent interfaces. Comprehensive tests will be done to assess the model's accuracy and dependability across numerous real-world scenarios and datasets, assuring its applicability and resilience. The insights collected are meant to guide future research in the dynamic field of emotion detection and contribute to the continual improvement of technology at the interface of emotion and HCI.

However, the research faces significant drawbacks. The variety in the quality and diversity of datasets utilized may influence the model's accuracy and dependability since variances in lighting, resolution, and facial angles might affect performance. Real-world applications may provide unanticipated obstacles, such as differences in user behavior and environmental elements that might disrupt the model's performance. Additionally, building and fine-tuning deep learning models need large computational resources, and restricted access to high-performance computers may hamper the research's scope and pace. Lastly, the use of face recognition technology involves ethical considerations, including privacy issues and potential misuse, needing careful study to ensure responsible development and deployment.

1.6 Report Organization

The suggested report is organized thoroughly to walk readers through the research steps, conclusions, and ramifications of employing deep learning models for emotion detection. Every chapter has a distinct function and adds to the document's overall coherence and depth.

Chapter 1: Introduction

The introductory chapter provides background on the significance of emotion detection, the application of deep learning models, and the need for a comparative analysis of different deep learning approaches. It outlines the research questions, objectives, and the overall framework of the study.

Chapter 2: Literature Review

The background chapter presents a detailed exploration of relevant literature and prior research. This section offers an in-depth understanding of the theoretical foundations, methodologies, and technological advancements related to emotion detection, facial expressions, and deep learning. It establishes the context for the current study and highlights gaps in existing knowledge.

Chapter 3: Methodology/ Requirement Analysis & Design Specification

The research methodology chapter outlines the approach taken to conduct the study. It provides insight into the research design, data collection methods, model architecture, and the rationale behind choosing specific methodologies. The chapter also discusses ethical considerations and any limitations that might impact the study.

Chapter 4: Implementation

This section outlines the detailed implementation process of the study, including data preparation, model training, and evaluation procedures used to enhance facial emotion recognition using advanced deep learning models.

Chapter 5: Result and Analysis

This chapter presents the experimental results of applying various deep learning models to emotion detection. Detailed analyses of the performance of models like Vision Transformer (ViT), VGG19, InceptionV3, Efficient-Net, and ResNet50 are provided, along with visual representations and statistical measures to validate the outcomes.

Chapter 6: Impact on Society, Environment and Sustainability

The impact chapter explores the broader implications of the research findings on society. It discusses how improved emotion detection methods can enhance user experiences, make interactions more natural and empathetic, and positively influence various applications such as virtual assistants, customer service bots, and healthcare diagnostics.

Chapter 7: Conclusion and Future Work

This final chapter serves as an overview of every aspect of the report. It summarizes the key findings, reiterates the conclusions drawn from the research, and offers recommendations for practical applications and further studies.

1.7 Summary

The research conducted the essential role of facial expressions in communication, especially in Human-Computer Interaction (HCI). It focuses on applying deep learning models to effectively detect emotions from looking at images. The main challenge addressed is the difficulty machines have in accurately understanding human emotions, which hampers seamless interaction between humans and computers. This study aims to develop emotionally intelligent interfaces by creating a method to detect six basic emotions: anger, neutrality, fear, happiness, sadness, and surprise. The scope of the research involves extensive tests to ensure that the model can capture complicated facial emotions in real-world circumstances. Additionally, the study goes into the historical and theoretical components of emotion detection, showing observations that might guide future research. Expected outcomes include a validated emotion detection model, a deeper understanding of how emotions are recognized, and advancements in HCI technology. This research tries to bridge the gap in present technology, improving how machines interpret and respond to human emotions.

CHAPTER 2

Literature Review

2.1 Overview

Emotion Detection uses facial images to classify human emotions anger, neutrality, happiness, sadness, disgust, and surprise This work addresses making use of the previously described deep learning approaches to how these emotions can be identified and displayed. The emphasis is on evaluating the relative performance of many models, therefore stressing their advantages and drawbacks in terms of face expression capture of subtleties. The work investigates the synergy among many deep learning models through their complementing behavior in the framework of emotion detection. Employing performance comparison, the study seeks to pinpoint the most effective methods for accurately determining emotions from facial expressions. Understanding the unique features of every model and how it applies in practical situations depends on this comparison study.

2.2 Related Works

In 2023, Gautam et al. [1] integrated CKplus and Jaffe datasets for emotion identification, leveraging feature extraction methods like HOG and SIFT mixed with Convolutional Neural Networks (CNN). Achieving excellent accuracy rates: HOG with CNN (78.48%), the study highlights the usefulness of the suggested technique, displaying better performance in emotion recognition compared to state-of-the-art models.

In 2022, Avanija et al. [2] utilized varied datasets, including CK+, JAFFE, and FER2013, applying approaches such as Viola-Jones + CNN with an accuracy of 88.56. Despite problems like generalization and high complexity, the article underlines the relevance of establishing optimum model combinations for increased FER performance in Human-Machine Interaction.

In 2022, Kanagaraju et al. [3] implemented a CNN-based VGG16 architecture trained on the FER-2013 dataset to recognize facial expressions in real-time with 64.52% accuracy. While the system categorizes emotions and provides video-based solutions for sadness, anger, and fear, there is room for improvement in accuracy, micro-expression identification is challenging, and emotion detection may be biased. The research

emphasizes the need for improved face expression recognition accuracy and ethics in intelligent human-computer interaction.

In 2017, Dachapally et al. [4] used two methods for emotion recognition: autoencoders for unique emotion representation and an 8-layer CNN. Trained on the JAFFE dataset and evaluated using the LFW (Labeled Faces in the Wild) dataset, our CNN outperforms state-of-the-art methods with JAFFE achieving 86.38% accuracy and LFW reaching 67.62%. The study suggests potential improvements for autoencoders.

In 2018, Syed et al. [5] extended the Cohn-Kanade dataset and a 32-image test set for emotion recognition. Fisher Face had 56% accuracy, especially with disdain, but the dual-classifier (Fisher Face plus HOG) equaled HOG's 81% accuracy in 20% less time. A combination of Viola-Jones, Harris corners, PCA, LDA, HOG, and SVM can recognize seven fundamental facial emotions, including disdain.

In 2018, Ko et al. [6] examined facial emotion recognition (FER), covering traditional and deep-learning-based FER techniques. The research delves into representative categories and algorithms, emphasizing the hybrid CNN + LSTM method's astounding 78.61% accuracy. However, problems exist, including the necessity for large-scale datasets and addressing the complexity of micro-expressions. The study serves as a comprehensive reference for both newbies and experienced researchers, providing insights into recent developments and probable future paths in FER.

In 2020, Jaiswal et al. [7] introduced an AI system for emotion detection through facial expressions, employing a three-step process: face detection, feature extraction, and emotion classification. The proposed method utilizes a convolutional neural network (CNN) architecture and achieves accuracies of 70.14% (FERC-2013 dataset).

In 2020, Ali et al. [8] employed facial images to analyze human emotions, focusing on Darwin's foundational work. Identifying seven universal emotions - neutral, angry, disgust, fear, happy, sad, and surprise - the research proposes an effective method using Convolutional Neural Network (CNN) in Keras to detect neutral, happy, sad, and surprise emotions. The CNN algorithm excels at tasks like object detection, face recognition, and image processing through deep convolutional neural network (DCNN) layers.

In 2020, Mehendale et al. [9] addressed current facial investigation issues, including unnecessary features due to retaining the background, causing confusion in CNN training. This study concentrates on displeasure anger, sad, happy, fear, and surprised expressions. The FERC algorithm aims to categorize images into these five emotion classes, achieving accuracies of 85% (Caltech faces), 78% (CMU database).

In 2020, Saxena et al. [10] analyzed over a hundred papers, focusing on methodologies and datasets for emotion detection. Facial expression recognition is compared using standard databases (JAFFE, CK+, Berlin Emotional Database). The study identifies seven basic emotions and compares detection methods. Notable results include Particle Swarm Optimization using Biogeography-Based Optimization for Emotion Recognition, Stationary Wavelet Transform (Facial Emotion Recognition), and varied accuracies for different methods: Artificial Neural Networks (73%), Support Vector Machines (86%), Hybrid Deep Neural Network (77%).

In 2021, Modi et al. primarily [11] focused on the challenges faced while transitioning from restricted laboratory environments to real-world scenarios. The text emphasizes the need to integrate Facial Expression Recognition (FER) technology into robotic systems, offering crucial insights for further advancements in this domain. In this study, Facial Emotion Recognition (FER) achieved the highest accuracy of 82.51% using Convolutional Neural Networks (CNN) and Transfer Learning. This performance surpassed the accuracy of Transfer Learning, which achieved 73.58%, during the 35th epoch.

In 2015, Gajarla et al. [12] focused on sentiment analysis and emotion prediction from Photo-sharing websites such as Flickr and social media platforms like Twitter. The goal is to predict emotions in images such as Love, Happiness, Violence, Fear, Sadness using fine-tuned convolutional neural networks. Results show accuracies of 67.8% (VGG-ImageNet), 68.7% (VGG-Places205), and 73% (ResNet-50).

In 2013, Alonso-Martin et al. [13] introduced a multimodal system for detecting user emotions for social robots integrated into the Robotics Dialog System (RDS). The system incorporates Gender and Emotion Facial Analysis (GEFA), utilizing third-party solutions SHORE and CERT. Emotion detection accuracies are reported as JRIP achieved an accuracy of 57.10%, while SHORE demonstrated a higher accuracy of 75.55%, and CERT achieved an accuracy of 62.66%.

In 2016, Mollahosseini et al. [14] provided a model with two convolutional layers and four Inception layers, outperforming state-of-the-art techniques and traditional networks, highlighting its potential for robust, generalized FER, especially in subject-independent and cross-database circumstances. Utilizing a deep neural network for the recognition of facial expressions (FER) covering seven datasets, it approximately corresponds with or slightly outperforms AlexNet in accuracy for most datasets.

In 2017, Madinah et al. [15] used a method that combines facial feature extraction involves the use of Harris corner key-points and Viola-Jones cascade object detectors. Utilizing principal component analysis, linear discriminant analysis, HOG feature extraction, and SVM, it classifies seven human facial expressions. The approach enables initial classification, achieving reasonable accuracy (81%) with a 20% faster runtime than using HOG exclusively.

In 2023, Mamieva et al. [16] provided a distinctive attention-based technique for multimodal emotion identification, incorporating face and voice information from IEMOCAP and CMU-MOSEI datasets. The proposed system is looked at on two datasets, finding that, on the IEMOCAP dataset, the weighted accuracy WA was 74.6%, and the F1 score was 66.1%. On the CMU-MOSEI dataset, the equivalent results were 80.7% and 73.7%.

In 2018, Turabzadeh et al. [17] presented a facial expression recognition system, initially tested on MATLAB, achieved promising results at 1 frame per second. To enhance real-time performance, it was implemented on an FPGA with a Digilent VmodCAM stereo camera module. The system was evaluated on a dataset of five users with 63,000 samples, resulting in an overall accuracy of 47.44%.

In 2017, Uddin et al. [18] suggested a system for recognizing facial expressions that integrates innovative approaches, such as LDRHP and LDSP, with KPCA, GDA, and CNN, based on depth information. With a remarkable identification rate of 86.25%, it surpasses conventional approaches such as LD2BP-HMM (75.83%) demonstrating its efficacy in tackling issues related to depth-based expression recognition.

In 2020, Joseph et al. [19] focused on facial emotion identification using geometry-based features. Fuzzy and discrete wavelet transform (DWT) methods yield enhanced images. TensorFlow is used to classify the extracted characteristics because of its pipelining capabilities and flexibility. The accuracy is obtained by testing using threedatabases:

expanded Cohn–Kanade (CK+), Oulu–CASIA NIR–VIS facial expression (Oulu–CASIA), and Karolinska Directed Emotional Faces (KDEF).

In 2020, Alreshidi et al. [20] used AdaBoost cascade classifiers for face detection and Neighborhood Difference Features (NDF) for feature extraction. The Random Forest Classifier-based emotion classification system outperforms five reference methods on the SFEW and RAF datasets. The framework's modularity, independence from gender and face skin color, and robustness to lighting and orientation fluctuations make it attractive for HCI.

2.3 Comparison between existing works

SL No	**Author Name**	**Used Algorithm**	**Best Accuracy with Algorithm**
1	Gautam et al. [1]	HOG, SIFT, CNN	HOG with CNN: 78.48%
2.	Avanija et al. [2]	Viola-Jones+CNN	Viola Jones + CNN: 88.56%
3.	Kanagaraju et al. [3]	CNN-based VGG16	CNN-based VGG16: 64.52%
4.	Dachapally et al. [4]	Convolutional Neural Networks	CNN: 86.38% (JAFFE dataset)
5.	Syed et al. [5]	Fisher Face, HOG, Viola-Jones, Harris corners, PCA, LDA, SVM	Dual-Classifier (Fisher Face + HOG): 81%
6.	Ko et al. [6]	Hybrid CNN + LSTM	Hybrid CNN + LSTM: 78.61%
7.	Jaiswal et al. [7]	Convolutional Neural Network (CNN)	CNN: 70.4% (FERC-2013)

8.	Mehendale et al. [9]	FERC	FERC: 85% (Caltech faces), 78% (CMU database)
9.	Saxena et al. [10]	ANN, SVM, Constructive feedforward neural networks, Adaboost	SVM 86%, ANN: 73%
10.	Modi et al. [11]	Convolutional Neural Networks (CNN) and Transfer Learning	CNN: 82.51%
11.	Gajarla et al. [12]	ResNet50, VGG16, and VGG-ImageNet	ResNet-50: 73%
12.	Alonso-Martin et al. [13]	Gene Expression Feature Analysis (Short Homozygosity Runs Evaluator, Classification Error Rate Testing), JRIP	SHORE: 75.55%
13.	Mollahosseini et al. [14]	Convolutional and Inception layers, AlexNet	Outperforms AlexNet in most datasets

2.4 Open Issues

Recent studies show that facial expression recognition research presents a broad range of difficulties as well as progress. The stated performance measures range greatly, from 64.52% to 88.56%, highlighting the challenge in obtaining consistent accuracy across several datasets and approaches. Techniques include conventional approaches such as HOG and Viola-Jones to cutting-edge deep learning models including CNNs, CNN with LSTM, and Transfer Learning. Every strategy displays various levels of success, therefore stressing the lack of a generally better solution. Results are strongly influenced by dataset properties; differences amongst datasets such as JAFFE, FERC, and NIST highlight the need for strong, universal models. Although hybrid models combining CNNs with other approaches generally show competitive performance, improving these hybrids for more general use remains a top priority. Further difficulties arise from computational efficiency and ethical issues like privacy and bias in dataset representation. Fair comparisons and promotion of repeatability depend on standardized standards and assessment criteria. Refine models for precisely capturing complex human

expressions in many circumstances by use of collaborative efforts between computer vision and psychology. Dealing with this complexity will be essential to improve face expression recognition toward more dependable, relevant, and morally sound uses.

2.5 Summary

In facial emotion recognition, CNNs excel at learning features from images, while RNNs and LSTMs capture sequential dependencies, making them ideal for time-series data. Hybrid models combining CNNs with LSTMs improve accuracy. Transfer learning, especially from ImageNet, enhances performance by fine-tuning pre-trained models. Attention mechanisms improve feature capture by focusing on specific facial regions. Traditional methods analyze facial landmarks, while ensemble methods combine predictions for robust results. Real-time implementation requires optimization for edge devices with GPUs and TPUs to ensure low latency. Recent studies show varying performance, indicating the challenge of consistent accuracy across datasets like JAFFE, FERC, and NIST. While hybrid models show competitive performance, improving their generalizability remains crucial. Computational efficiency, privacy, and bias in datasets are key challenges. Standardized evaluation criteria and collaborative efforts between computer vision and psychology are essential for FER.

CHAPTER 3

Methodology/Requirement Analysis & Design Specification

3.1 Overview

Our study project has a distinct methodology, detailed in this section. The aim is to discover successful techniques for solving emotion recognition problems using advanced deep learning methods. We applied several deep learning models, including Vision Transformer (ViT), Efficient-Net, VGG19, InceptionV3, and ResNet50, chosen for their capabilities in handling complex image recognition tasks.

Before running any models, it was important to have a well-prepared dataset. We used a large dataset of facial images, ensuring meticulous preprocessing and augmentation to improve the training process, which is vital for getting accurate results. We trained each model on the preprocessed dataset to adapt them to our specific goal of facial emotion recognition. This involved adjusting hyperparameters and using optimization methods to improve model performance and reduce overfitting. To determine the models, we used performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices. These metrics gave a comprehensive assessment of each model's ability to correctly recognize emotions from facial expressions. The methodology part explains how various components of our approach work together to produce excellent results, including graphical representations and mathematical equations to aid in understanding the entire process. A clear and detailed workflow chart shows the step-by-step process, making it easier for readers to grasp the methodology quickly.

By preparing the dataset carefully, training and fine-tuning advanced models, and using robust evaluation metrics, we aimed to achieve high accuracy and reliability in emotion detection. The detailed workflow and justifications ensure transparency and reproducibility, important for future study and practical applications.

For further validation, we are aiming to validate our dataset by a licensed psychologist. By doing so, we can confirm and compare the results gained by the models as opposed to the psychologist's classification.

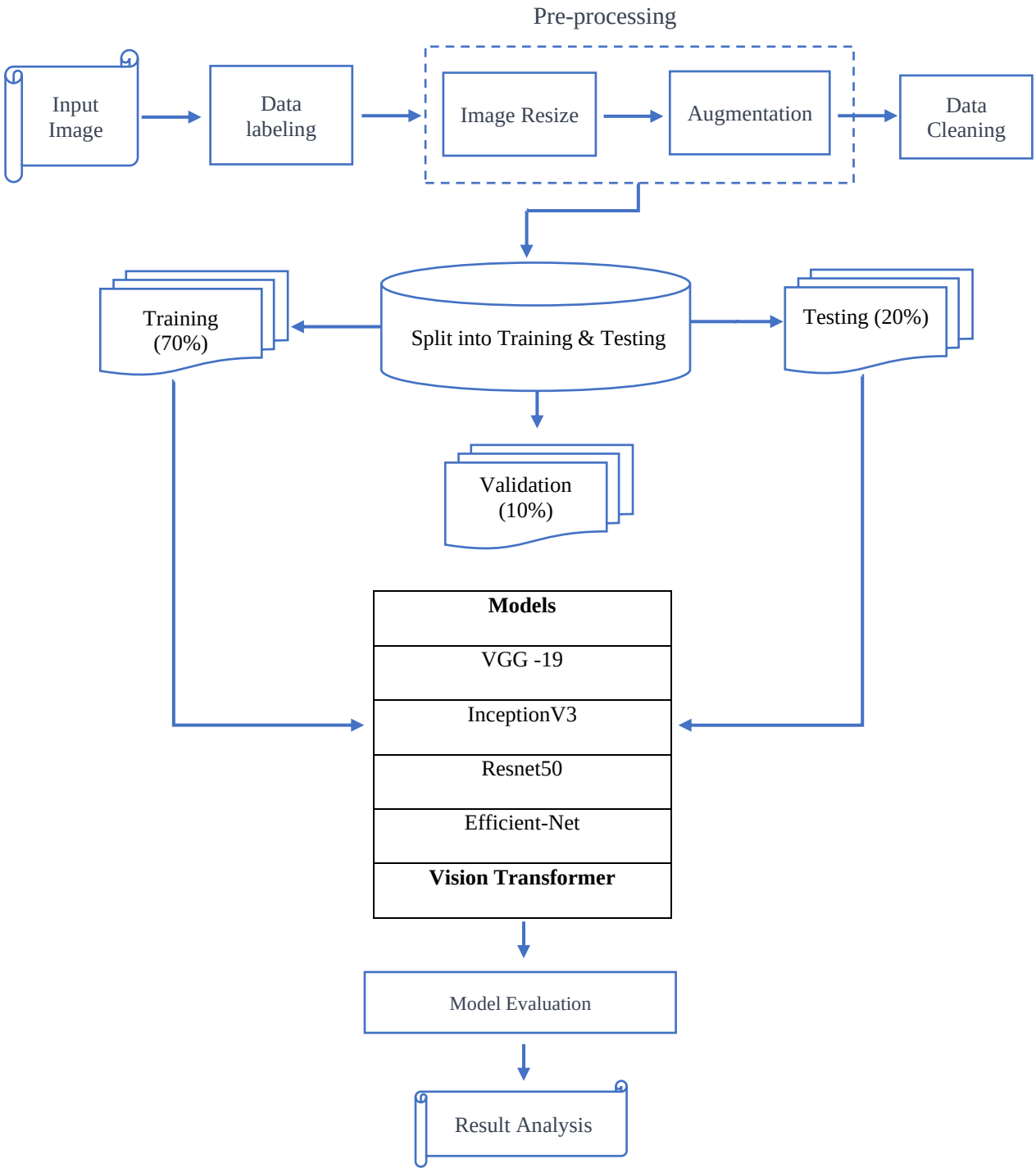


Figure 3.1: Work Process for our Methodology

3.2 Data Collection Procedure/Dataset Utilized

3.2.1 Data Collection

During the data gathering phase, we used a mobile camera to collect 1500 raw images from 250 individuals, representing six classes: Happy (249 images), Sad (242 images), Neutral (240 images), Angry (228 images), Surprise (239 images), and Disgust (224 images). These images were captured under various lighting conditions and backgrounds to ensure the model could handle real-world variations. In the data augmentation phase, we enriched our dataset by creating new training images through random transformations of the existing ones, including rotations, translations, shears, magnifications, and reflections. These modifications aimed to enhance the model's robustness and its ability to generalize to new data. Subsequently, when we advanced to the data augmentation phases, we commenced on the process of enriching our dataset by producing new training photos via random alterations of the current ones. These alterations comprised a variety of transformations, including rotations, translations, and shears among others. The purpose underlying these modifications was to boost the model's robustness to variances in input data, hence increasing its capacity to generalize to unknown data.

Figure 3.2.1: Sample Data set

Table 3.2.1: Distribution of the dataset before augmentation

Class Name	Quantity
Happy	249
Sad	242
Neutral	240
Angry	228
Surprise	239
Disgust	224

3.2.2 Data Preprocessing

Data preprocessing is a crucial stage in developing deep learning models, as it directly affects the model's ability to learn and generalize from the input data. For our image dataset, we applied several preprocessing techniques:

1. Resize: In order to maintain consistency and enable deep learning models like Vision Transformer (ViT), Efficient-Net, VGG19, ResNet50, and InceptionV3 to learn consistent features from each picture, we downsized the images from 1080x1080 pixels to 224x224 pixels.

Figure 3.2.2.1: Sample Data set (Original vs Resized)

2. Noise reduction: To reduce noise in the photos and prevent it from interfering with the training process, Gaussian smoothing was used.

Figure 3.2.2.2: Sample Data set (Original vs Smoothed)

3. Data normalization: To scale the pixel values to the range [0, 1], the pixel values of each picture were divided by 255.

The resilience and generalization capacities of the deep learning models were improved by these preprocessing measures, which also improved the quality and variety of the dataset.

3.2.3 Data Augmentation

Data augmentation with the Augmentor library increased the size and diversity of our dataset, ensuring unique variations of the original images without losing essential information. The following augmentation techniques were applied: rotation, zoom, horizontal flip, and vertical flip. This process preserves the label data while enhancing the dataset, thereby improving the model's generalization and resilience to unknown inputs. To meet our training data objectives, we utilized brightness, contrast, scaling, flipping, and rotation. Specifically, data augmentation included random rotations, distortion, shearing, vertical and horizontal flips, and changes in brightness. Each original image was augmented 10 times, creating a more heterogeneous and robust dataset. These augmentation steps were essential for improving the performance and generalization capabilities of our deep learning models.

Table 3.2.3: Distribution of the dataset After augmentation

Class Name	Quantity
Happy	2000
Sad	2000
Neutral	2000
Angry	2000
Surprise	2000
Disgust	2000

3.2.4 Data Splitting

We created a dataset with a total of 12,000 images, consisting of 2,000 images per class. For model training, 70% (8,400 images) were used. Additionally, 20% (2,400 images) were reserved for testing to assess model performance. The remaining 10% (1,200 images) were set aside for validation to ensure the model performs well with new, unseen data.

Table 3.2.4: The distribution of data among the training, testing, and validation sets.

Class Name	Train	Test	Validation	Total
Happy	1400	400	200	2000
Sad	1400	400	200	2000
Neutral	1400	400	200	2000
Angry	1400	400	200	2000
Surprise	1400	400	200	2000
Disgust	1400	400	200	2000

3.3 Proposed Methodology/ System Design

3.3.1 VGG19

VGG-19 is a deep CNN pre-trained for complex classification tasks, featuring 19 layers. It includes convolutional layers for feature extraction, max-pooling for size reduction, and fully connected layers for classification. The architecture uses 3x3 convolution filters and ReLU activation to address the vanishing gradient issue, with max-pooling layers down-sampling the output. Training involves forward and backward propagation to optimize weights, with batch size influencing the number of samples per epoch. Pre-trained layers are often frozen, with a shallow network added for customized tasks. Dropout layers prevent overfitting by randomly deactivating neurons during training. VGG-19's deep architecture and regularization techniques enhance accuracy, making it effective for various image classification tasks [26]. Here's a concise step-by-step explanation of the VGG19 processes images, including relevant equations,

Convolutional Layer: F_i = ReLU ($W_i * X * b_i$)

Max Pooling Layer: P_i = maxpool (F_i)

Fully Connected Layer: D_j = ReLU $(W_j \cdot P + b_j)$

Dropout Layer: D'_j = dropout (D_j, p)

Loss Function: $L = -\sum_{c=1}^{C} y_c * log_{(p_c)}$

Softmax Classifier: $\bar{y}$ = softmax ($W_{out} \cdot D + b_{out}$)

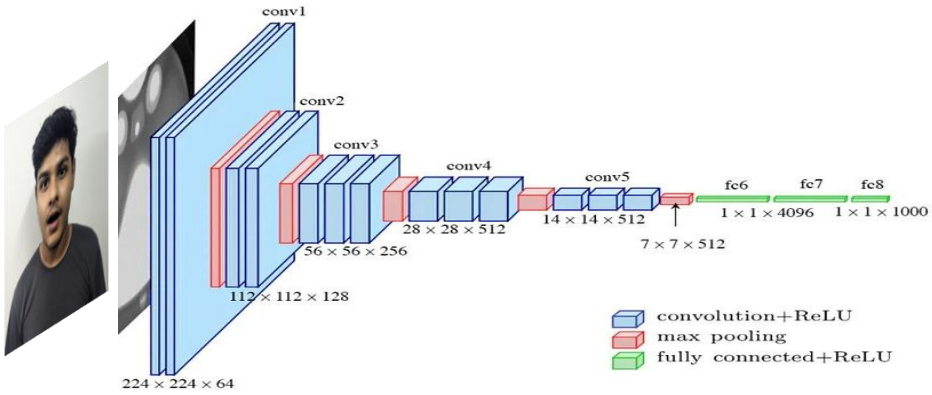


Figure 3.3.1: VGG-19 Architecture [22]

3.3.2 InceptionV3

InceptionV3 is a CNN used for image recognition tasks, featuring components like AvgPool, MaxPool, Convolution, Concat Layer, Fully Connected layer, Dropout, and Softmax. Convolution layers extract features, pooling layers reduce complexity, and concatenation merges tensors. Dropout mitigates overfitting, while fully connected layers handle classification. The softmax function provides final outputs. InceptionV3 excels in computer vision tasks like detection and classification and is often used for transfer learning, optimizing pre-trained weights for specific tasks. Its uniform 3x3 filters, consistent stride, and equal padding contribute to its simplicity and efficiency [27]. Here's a concise step-by-step explanation of how the InceptionV3 processes images, including relevant equations,

Convolutional Layer: F_i = ReLU ($W_i * X * b_i$)

Max Pooling Layer: P_{max} = maxpool (F_i)

Average Pooling Layer: P_{avg} = avgPool (F_i)

Concatenation Layer: C = concat (P_{max}, P_{avg},...)

Fully Connected Layer: D_j = ReLU $(W_j \cdot C + b_j)$

Dropout Layer: D'_j = dropout (D_j, p)

Softmax Classifier: $\bar{y}$ = softmax ($W_{out} \cdot D$ + b_{out})

Loss Function: L = - $\sum_{c=1}^{C} y_c * log_{(p_c)}$

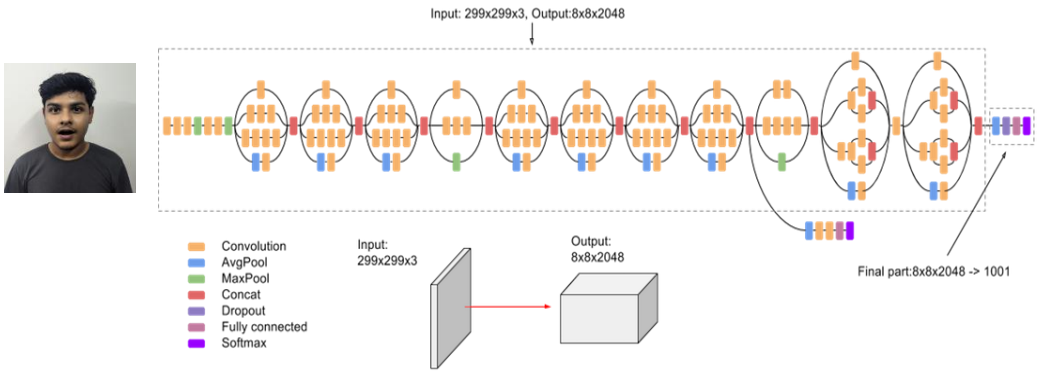


Figure 3.3.2: InceptionV3 Architecture [23]

3.3.3 Resnet50

ResNet-50 is a deep CNN for image classification that uses residual learning to address the degradation problem in deep networks. Its architecture consists of 50 layers, including 49 convolutional layers and 1 fully connected layer. The core component is the residual block with 1x1, 3x3, and another 1x1 convolution layers, featuring identity shortcut connections to mitigate the vanishing gradient problem. The architecture begins with an initial convolutional layer and max pooling, followed by four stages of convolutional layers with residual blocks, each stage having more filters. Batch normalization and ReLU activation are used after every convolutional layer. The network concludes with average pooling, a fully connected layer, and a SoftMax activation function for classification. ResNet-50's design allows for effective training of deep networks, making it a reliable choice for image classification and other computer vision tasks, even though newer models have built upon its concepts [28]. Here's a concise step-by-step explanation of how the Resnet50 processes images, including relevant equations,

Residual Block: $y = F(x, \{W_i\}) + x$

Convolutional Layer: $F_i = \text{ReLU}(\text{BatchNorm}(W_i * X * b_i))$

Identity Shortcut Connection: $y = F(x, \{W_i\}) + x$

1x1 Convolution: $F_{1x1} = \text{ReLU}(\text{BatchNorm}(W_{1x1} * X * b_{1x1}))$

3x3 Convolution: $F_{3x3} = \text{ReLU}(\text{BatchNorm}(W_{3x3} * F_{1x1} * b_{3x3}))$

Residual Function: $F(x, \{W_i\}) = \text{ReLU}(\text{BatchNorm}(W_{1x1.2} * \text{ReLU}(\text{BatchNorm}(W_{3x3} * \text{ReLU}(\text{BatchNorm}(W_{1x1,1} * X + b_{1x1,2})) + b_{3x3})) + b_{1x1,2}))$

Output of Residual Block: $y = F(x, \{W_i\}) + x$

Final Fully Connected Layer: $\bar{y} = \text{softmax}(W_{fc} * \text{AvgPool}(y) + b_{fc})$

Loss Function: $L = -\sum_{c=1}^{C} y_c * log_{(pc)}$

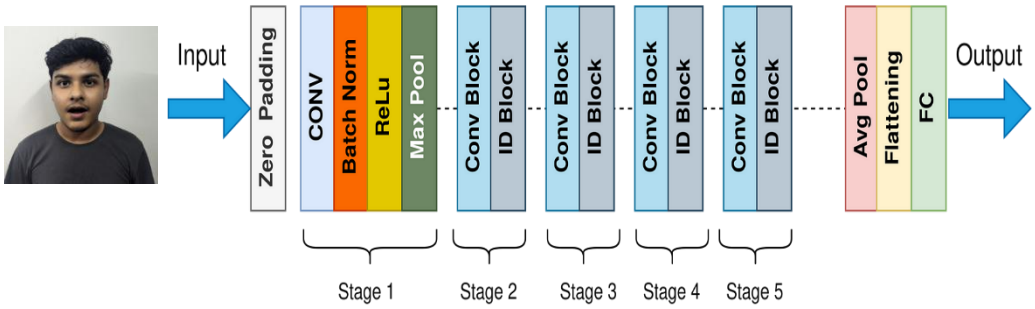


Figure 3.3.3: ResNet50 Architecture [24]

3.3.4 Efficient-Net

Efficient-Net is a CNN designed for excellent accuracy with low computing costs, ideal for low-cost devices. It uses a scaling method to adjust model sizes efficiently, making it effective for image recognition. The architecture starts with a 224x224 pixel input image, followed by a 3x3 convolutional layer, several MBConv layers, and ends with a 7x7 resolution feature map with 320 channels. The Efficient-Net family (B0 to B7) scales in depth, width, and resolution to enhance accuracy and complexity. Performance can be further improved with techniques like augmentation and semi-supervised learning [29]. Here's a concise step-by-step explanation of how the Efficient-Net processes images, including relevant equations,

Compound Scaling: $d = \alpha^{\emptyset},\ \omega = \beta^{\emptyset},\ r = \gamma^{\emptyset}$

MBConv Block: $y = x$ + Conv (SE (DepthwiseConv (Conv (x))))

Convolutional Layer: y = ReLU6 (BatchNorm ($W * x + b$))

Depthwise Convolution: $y_i = \sum_k W_{ik} * x_k + b_i$

Squeeze-and-Excitation: s = σ (W_2 * ReLU (W_1 *GlobalAvgPool(x)))

$$Y = s * x$$

Final Fully Connected Layer: $\bar{y}$ = softmax ($W_{fc} *$ AvgPool (y) + b_{fc})

Loss Function: $L = -\sum_{c=1}^{C} y_c * log_{(pc)}$

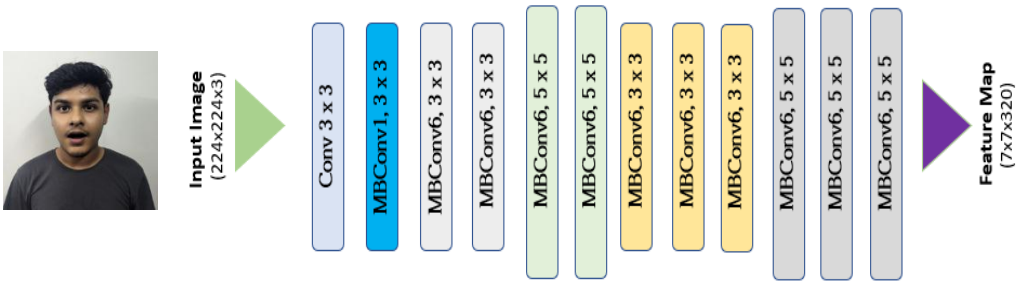


Figure 3.3.4: Efficient-Net Architecture [25]

3.3.5 Vision Transformer (ViT)

The Vision Transformer (ViT) divides a 224x224 image into 196 patches of 16x16 pixels each. Each patch is converted into a 768-dimensional vector, forming a 196x768 matrix. An additional "class token" vector is added, creating a 197x768 matrix. This matrix is processed by the encoder, consisting of Multi-Head Attention, normalization, and linear layers, to learn the relationships between patches and their classes. The updated class token is then converted into an output vector by the pooling block, which generates class predictions. ViT treats image patches like words in a language, using the Transformer's self-attention mechanism to capture long-range relationships and perform image classification, extending the Transformer's capabilities from text to visual data [30]. Here's the detailed step-by-step explanation of how the Vision Transformer (ViT) processes images, including equations where applicable:

Patch Embedding:

$$\text{Patch Embedding}(x) = \text{Flatten}(\text{Conv2D}(x, \text{filter size} = 16, \text{stride} = 16))$$

$$E = [E_1, E_2, \ldots E_{196}], E_i \in \mathbb{R}^{768}$$

Class Token Addition: $Z_0 = [E_{class}; E_1; E_2; \ldots; E_{196}], Z_0 \in \mathbb{R}^{197*768}$

Position Embedding: $Z_0 = Z_0 + P, P\ \mathbb{R}^{197*768}$

Multi-Head Attention: $\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) * V$

$$Q = ZW_Q, K = ZW_K, V = 2W_V$$

Layer Normalization: LayerNorm (x) $= \frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}}$

Feed Forward Network (FFN): FFN(x) = ReLU ($xW_1 + b_1$) * $W_2 + b_2$

Transformer Encode Block: $Z' =$ MultiHeadAttention (LayerNorm(Z)) + Z

$$Z_{out} = \text{FFN (LayerNorm } (Z')) + Z'$$

Output Layer: y = Softmax (Z_{class})

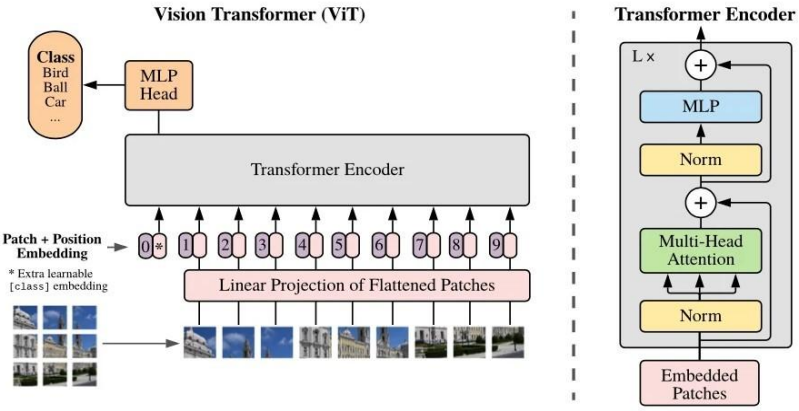


Figure 3.3.5.1: Vision Transformer Architecture [21]

3.3.6 Proposed Model

For our proposed work, the Vision Transformer (ViT) is particularly useful. We used this deep learning model to enhance the performance. The Vision Transformer (ViT) originates from the Transformer model, initially designed for natural language processing. Recognizing the potential of Transformers beyond text, researchers extended this architecture to the domain of computer vision. This model applies the same principles of self-attention and sequence modeling to images by treating image patches as tokens, similar to words in a sentence. This innovative approach demonstrated that Transformers could achieve competitive performance on image classification tasks, challenging the dominance of traditional convolutional neural networks (CNNs) in computer vision.

Below, we present our model structure with a graphical representation.

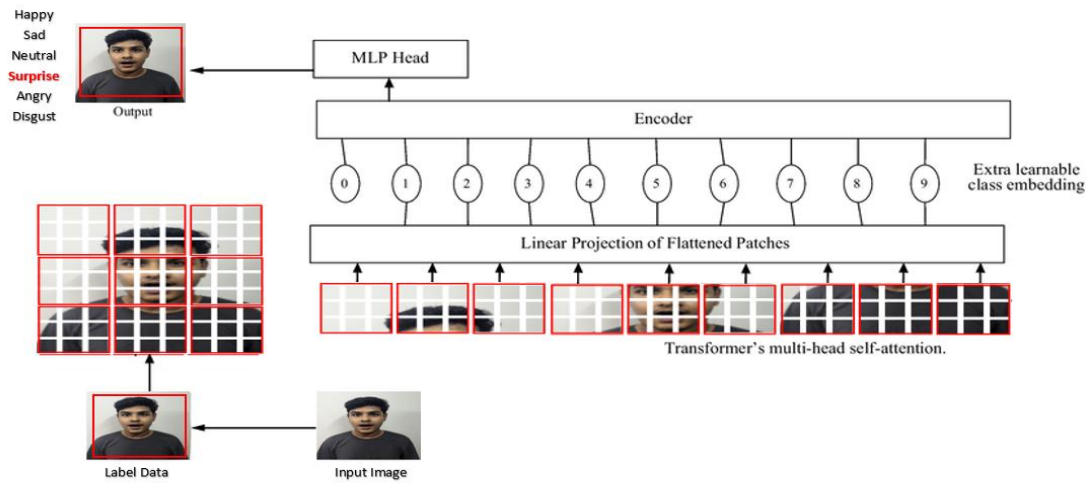


Figure 3.3.6.1: Architecture of our Proposed Method

3.4 Hardware/ Software Requirements

Automated Facial Expression Recognition research has certain hardware and software requirements. High-performance GPUs are required to accelerate both training and inference operations, especially for complex models like Vision Transformer and Efficient-Net. It is critical to use well-known deep learning frameworks such as TensorFlow or PyTorch for building models, as well as image processing tools like OpenCV or PIL for dataset preparation. This organized method assures that automated facial expression recognition systems are reliable and applicable in a variety of real-world circumstances

Hardware:

- Primary System:
  a) Processor: Intel Core i5 11th generation
  b) RAM: 16GB
  c) Storage: 512 GB SSD
  d) GPU: RTX 3050
- Cloud Computing: Google Colab with GPU access

Software:

- Operating System: Windows 11
- Programming Language: Python 3.7
- Deep Learning Framework: TensorFlow (1.15.2)
- Additional Libraries: NumPy, Pandas, ImageNet.

3.5 Project Management and Financial Analysis

3.5.1 Project management:

- **Research objectives:**
  A. Develop a model to identify human emotions from images accurately.
  B. To improve human-computer interactions by accurately identifying and responding to emotions, making interactions more authentic and sympathetic.
  C. To contribute to mental health monitoring by accurately classifying emotions, which is crucial for assessing well-being.
  D. To guide future researchers on this topic.

- **Research Timeline:**
  A. Phase 1: Planning and research (1-3 weeks)
  B. Phase 2: Review papers (2-4 weeks)
  C. Phase 3: Selection of most promising models (1 week)
  D. Phase 4: Dataset gathering (4-5 weeks)
  E. Phase 5: Building selected models (2-3 weeks)
  F. Phase 6: Training and testing (1-2 weeks)
  G. Phase 7: Getting human validation from a psychologist (1-2 weeks)
  H. Phase 8: Reviewing results and accuracy (1-2 weeks)

- **Resource Planning:**

A. Equipment and Tools:

- High-performance computer.
- Strong processor: Intel Core i5 11th gen.
- Graphics processing unit (GPU): RTX 3050

B. Software and Technologies:

- Python
- Google Colab.
- IDLE, Visual Studio Code, PyCharm
- Database storage: Google Drive, SSD (512 GB)
- Browser: Google Chrome, Opera, Firefox

C. Data and Content:

- **Images:** Provided by the participants through Google Drive and Google forum.
- **Report documentation:** Stored in Google Drive.

D. Training and Skill Development:

- Ensure that the development team has the necessary training and skills in deep learning model development and data preprocessing.
- Gain additional training on specific technologies, tools and models as needed.

E. Communication Plan:

- **Stakeholder Meetings:**
  Purpose: Evaluate the dataset by a licensed psychologist for authentication and validation.
  Participants: Team members, and stakeholders.

- **Change Control Meetings:**
  Purpose: Discuss and approve any changes to the project scope or requirements.
  Participants: Supervisor and team members.
  Frequency: As needed when change requests arise.

3.5.2 Financial Analysis:

Table 3.5.2: Estimated Cost for Research

SN	Components	Estimated Cost (BDT)
01.	Visiting Stakeholders	2000-3000
02.	Software and Tools	1000-2000
03.	Data Collection and Processing	1000-1500
04.	Documentation and Report Writing	200-500
05.	Miscellaneous (e.g., licenses, tools, devices)	500-1000
06.	Contingency (10% of total)	1000-1500
	Total Estimated Cost	6500-8000

3.6 Summary

This study aims to conduct a rigorous statistical analysis to determine the best-performing deep learning model for diagnosing Facial Expression. The dataset consists of 12,000 Facial images, evenly distributed across six classes, with each class containing 2,000 images. Initially, 70% of the total images (8,400) were allocated for training our computer model. This training phase ensures that the model learns effectively from a substantial portion of the dataset. For unbiased evaluation of model performance, 20% of the images (2,400) were reserved for testing. This separate test set allows us to assess how well the trained model generalizes to new, unseen data. Additionally, 10% of the total images (1,200) were set aside for validation. This validation set plays a crucial role in ensuring that the model performs consistently well across different subsets of data, guarding against overfitting and providing confidence in its real-world applicability. This assessment of the deep learning models and facilitates informed judgments on their usefulness in recognizing problems.

CHAPTER 4
Implementation

4.1 Overview

In this research, we start by loading the datasets and separating them into training and testing arrays. Using tensorflow.keras.utils, we reshape and translate the image labels of both the training and testing data into categorical values. TensorFlow and PyTorch create and train models, while OpenCV and PIL prepare datasets for augmentation, preprocessing, and normalization. Data augmentation improves training data diversity and robustness, while rigorous dataset curation assures face emotion portrayal. Vision Transformer (ViT), VGG19, InceptionV3, Efficient-Net, and ResNet50 are tested using hyperparameter modifying and normalization to optimize performance and reduce overfitting.

Detailed statistical analysis and visualization tools reveal each model's strengths and limitations in face expression recognition tasks. Accuracy and precision are essential indicators. Ethical standards, including patient data protection and informed consent, are upheld throughout the study to guarantee compliance with ethical principles. This experimental setting ensures a systematic and complete approach to investigating deep learning models for emotion detection, bringing vital insights and progress in automated facial expression recognition.

4.2 Train Model/ Prototype Design

The training process involves feeding the preprocessed images into a sequential model, which consists of two primary components: the convolutional network and the dense layer network. The model architecture includes three convolutional layers, each paired with a max-pooling layer to reduce dimensionality and capture important features. The output from the convolutional network is flattened and passed through the dense layers. Dropout is applied at the end of the first dense layer to prevent overfitting. Finally, the softmax classifier outputs the categorized images, translating the detected features into specific emotion labels. The training process shows promising accuracies for both the training and testing datasets, indicating that the models have been effectively trained.

4.3 System Testing/ Model Evaluation

Post-training, the model undergoes rigorous testing to evaluate its computational performance. Using the evaluate function, we test the model on the testing dataset to determine its accuracy and robustness in recognizing emotions. The evaluation reveals that the models, including ViT and others, are capable of recognizing emotions with high accuracy. The proposed approach is compared to existing models to highlight its efficacy. The train and test datasets provide sufficient data to recognize emotions accurately, confirming that our experimental procedures were successful. This detailed evaluation demonstrates that the models are well-trained and capable of effective emotion detection, contributing valuable insights to the field of facial emotion recognition. Several classification model performance criteria are used to assess deep-learning base models. Evaluations include ACC, precision, recall, F1-score, and confusion matrix (CM).

Accuracy: Accuracy measures how often the model correctly classifies images. It is calculated by dividing the sum of true positives and true negatives by the total number of samples. The formula is:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Here, TP is the number of correct positive predictions, TN is the number of correct negative predictions, FP is the number of incorrect positive predictions, and FN is the number of incorrect negative predictions.

Recall: Also known as sensitivity, recall quantifies the number of correctly identified positive predicted events. Recall is a useful statistic in cases where FN outperforms FP. Since recall and precision are harmonic means, a thorough comprehension of these two metrics can be obtained from the F1 score. When Precision and Recall are equal, it performs best

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Precision: Precision indicates the proportion of correct positive predictions out of all positive predictions made by the model. It is crucial when the cost of false positives is high. Precision is calculated using the formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

F-1 Score: This recall-precision measure of classification accuracy is called the F-1 score. The F1-score provides a comprehensive view since it is the harmonic mean of Precision and Recall. When precision and recall are equal, they are ideal.

$$\text{F1-score} = \frac{2*(\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

Confusion Matrix (CM): The confusion matrix is a table that summarizes the performance of a classification algorithm. It shows the counts of TP, FP, TN, and FN. The CM helps visualize how well the model distinguishes between different classes and provides insights into specific areas where the model may be making errors.

4.4 Summary

This research aims to enhance facial emotion recognition through advanced deep learning models. The process begins with loading and splitting datasets into training and testing arrays, followed by reshaping and categorizing image labels. The images are expanded and normalized for consistency. Models like Vision Transformer (ViT), Efficient-Net, VGG19, InceptionV3, and ResNet50 extract crucial features to improve detection accuracy. The training involves a sequential model with convolutional and dense layers, using max-pooling and dropout techniques to prevent overfitting. The softmax classifier categorizes the emotions. The model is then tested, showing high accuracy in emotion recognition. Comparisons with existing models confirm the effectiveness of the approach, highlighting the robust training and testing procedures and the contribution to the field of facial emotion recognition.

CHAPTER 5

Result and Analysis

5.1 Overview

This study focuses on developing a robust method for detecting six fundamental emotions Happy, Sad, Neutral, Angry, Surprise, and Disgust through facial expressions using deep learning models. By leveraging advanced technology and existing research, the objective is to enhance human-computer interaction by enabling computers to accurately recognize and respond to the diverse spectrum of emotions conveyed through facial expressions. Various pre-trained deep learning models including Vision Transformer, VGG-19, InceptionV3, Efficient-Net, and ResNet50 were employed, each demonstrating varying levels of accuracy. ViT emerged as the top performer with an accuracy of 82.96%, followed closely by Efficient-Net at 82.36%, ResNet50 at 80.87%, InceptionV3 at 79%, and VGG-19 at 78.22%. Based on these results, Vision Transformer (ViT) is recommended as the superior model for emotion recognition due to its precise accuracy and efficient design, promising exceptional performance and user experience. This advancement holds potential beyond technical achievements, offering avenues for monitoring mental well-being, enhancing psychological assessments, and improving human-computer interaction through real-time emotion detection in various devices.

5.2 Experimental/ Simulation Result

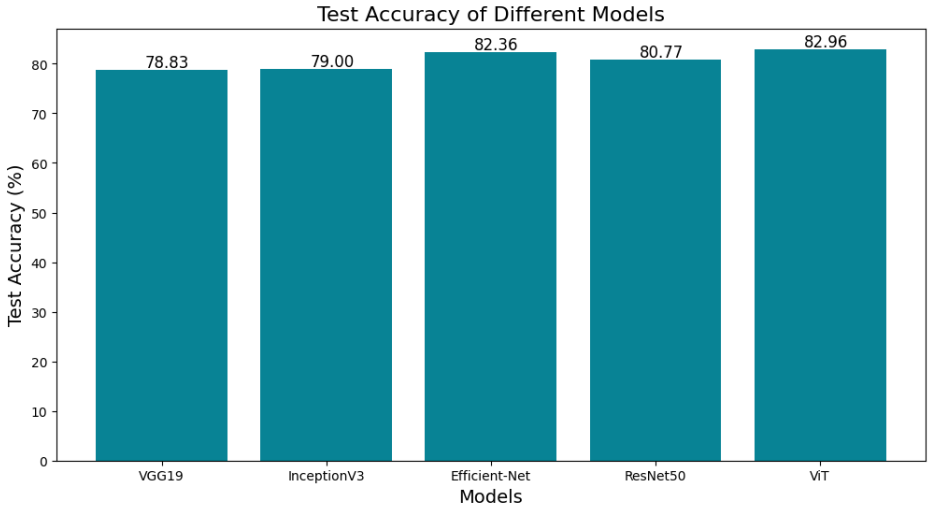


Figure 5.2.1: Accuracy Score of Different Models

Here, Table 5.2.1 displays the train, test, and validation accuracy and losses for our proposed model Vision Transformer (ViT) and the other four transfer learning models.

Table 5.2.1: Result of Vision Transformer and other Four Transfer Learning Models

Model	Train Accuracy	Train Loss	Validation Accuracy	Validation Loss	Test Accuracy
VGG19	77.23	0.6516	78.83	0.6099	78.83
InceptionV3	77.33	0.6295	77.92	0.6106	79.00
Efficient-Net	97.95	0.0846	81.55	0.7067	82.36
ResNet50	78.86	0.9997	80.87	0.9589	80.77
Vi-T	**91.58**	**0.2395**	**82.01**	**0.5945**	**82.96**

The comparative analysis of several pre-trained deep learning models for emotion identification shows diverse performance characteristics. Efficient-Net obtained remarkable train accuracy of 97.95% but exhibited evidence of overfitting with a considerable decline in validation accuracy of 81.55% and test accuracy of 82.36%. In contrast, Vision Transformer (ViT) demonstrated strong performance with a train accuracy of 91.58% and a train loss of 0.2395, reflecting efficient learning. It achieved the highest validation accuracy of 82.01% and the lowest validation loss of 0.5945 among all models. Its test accuracy was the highest at 82.96%, signifying robust generalization and superior performance in recognizing emotions across varied. While VGG19, InceptionV3, and ResNet50 displayed reasonable performance, ViT's strong generalization and better accuracy make it the recommended choice for emotion identification in this investigation.

Table 5.2.2: Compare the Precision, Recall, and F1-score of Vision Transformer and other Four Transfer Learning Models

Model	Types	Precision	Recall	F1-score	Accuracy
VGG19	Macro average Weighted average	81	79	79	79
InceptionV3	Macro average Weighted average	81	79	79	79
Efficient-Net	Macro average Weighted average	84	83	84	83
ResNet50	Macro average Weighted average	83	81	81	81
Vi-T	**Macro average** **Weighted average**	**83**	**83**	**83**	**83**

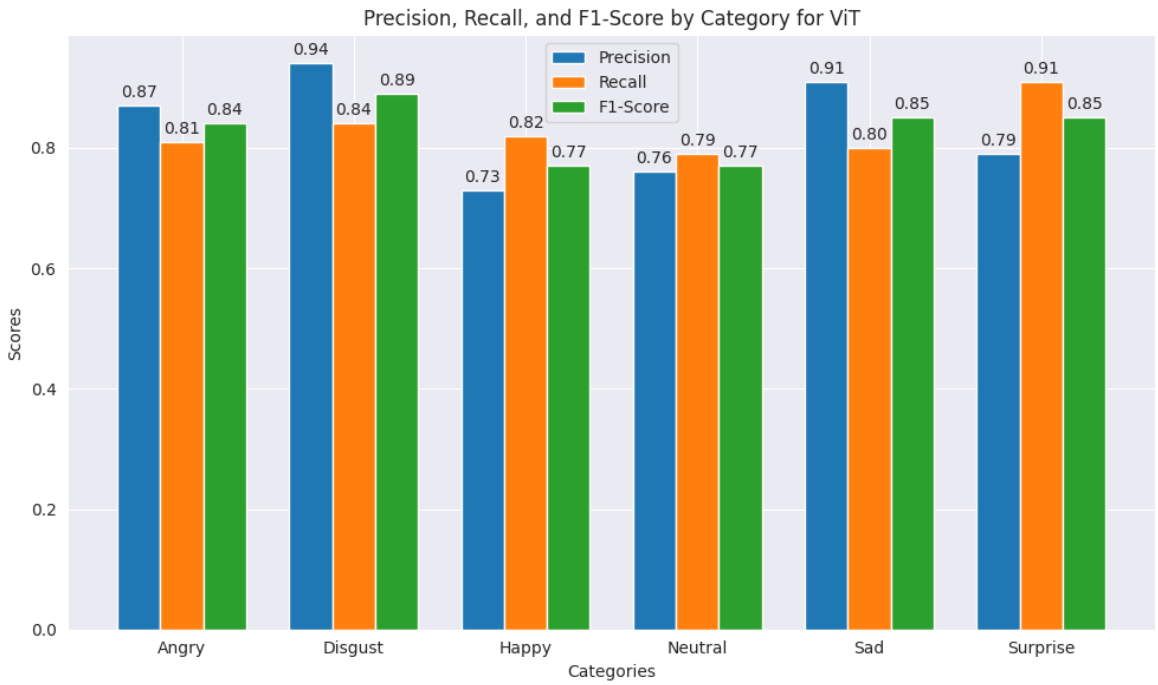
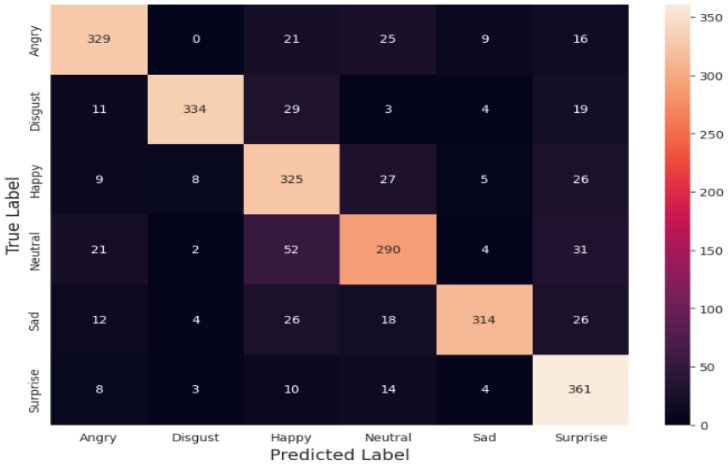


Figure 5.2.2: Evaluate the compared result of ViT for each class

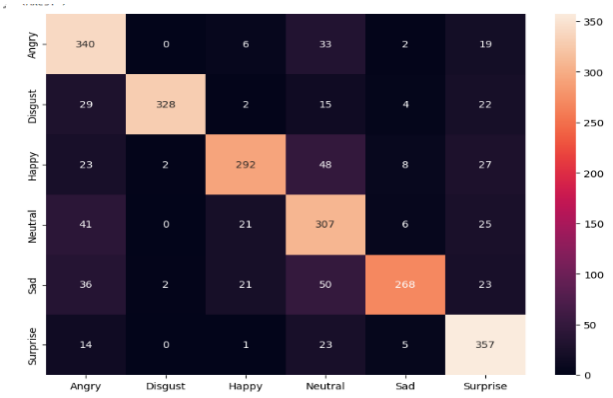
The comparative analysis of deep learning models for emotion recognition highlights their performance based on precision, recall, F1-score, and accuracy. Both VGG19 and InceptionV3 achieved a precision, recall, and F1-score of 79, with an accuracy of 79%. Efficient-Net outperformed these models with a precision of 84, recall of 83, and an F1-score of 84, achieving an accuracy of 83%. ResNet50 showed slightly lower performance with a precision of 83, recall of 81, and an F1-score of 81, maintaining an accuracy of 81%. The Vision Transformer (ViT) demonstrated the best balance with precision, recall, and F1-score all at 83, and an overall accuracy of 83%, making it the most consistent and reliable model for emotion recognition in this study.

5.3 Performance/ Comparative Analysis

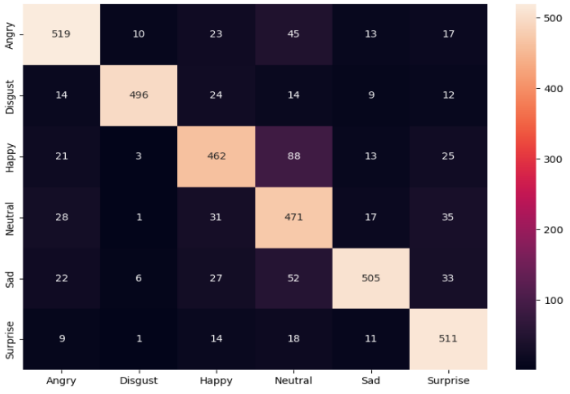
The comparative analysis of several pre-trained deep learning models for emotion identification revealed diverse performance characteristics. Vision Transformer (ViT) performed robustly with a training accuracy of 91.58% and a training loss of 0.2395. It achieved the highest validation accuracy of 82.01%, the lowest validation loss of 0.5945, and the highest test accuracy of 82.96%. ViT also had balanced precision, recall, and F1-scores, all at 83%, with an overall accuracy of 83%. In contrast, EfficientNet achieved a high training accuracy of 97.95% but showed signs of overfitting, with validation and test accuracies dropping to 81.55% and 82.36%, respectively. It had a precision of 84%, recall of 83%, F1-score of 84%, and an overall accuracy of 83%. VGG19 and InceptionV3 both had precision, recall, and F1-scores of 79%, and an accuracy of 79%. However, they struggled with specific emotion classifications, leading to notable misclassifications. ResNet50 showed acceptable performance with a precision of 83%, recall of 81%, F1-score of 81%, and an accuracy of 81%, but it had room for improvement in consistency. Overall, the ViT model emerged as the most reliable for emotion identification, with the best generalization capabilities and balanced performance across all metrics. EfficientNet, despite its high accuracy, struggled with overfitting and misclassifications. VGG19, InceptionV3, and ResNet50 displayed reasonable performance but were less consistent compared to ViT.



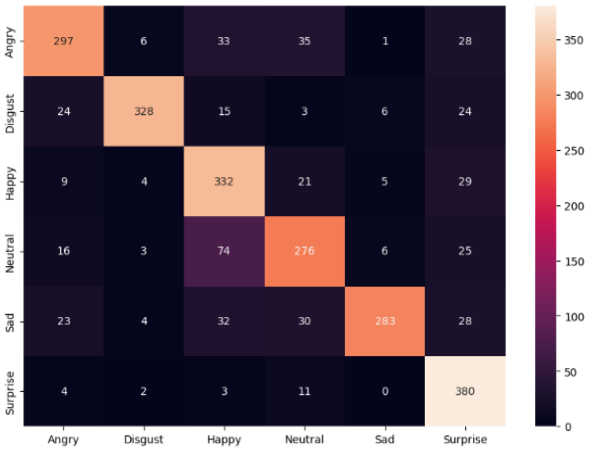
(a) Vision Transformer



(b) VGG-19



(c) Efficient-Net



(d) InceptionV3

(e) ResNet50

Figure 5.3.1: confusion matrix of the classification results of all models (a) Vision Transformer, (b) VGG19, (c) Efficient-Net, (d) InceptionV3, and (e) ResNet50

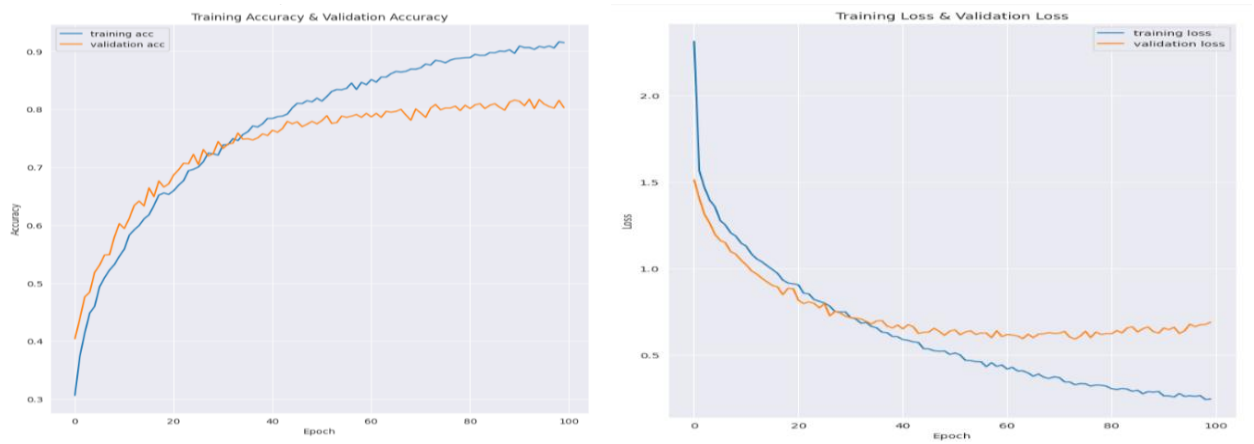


Figure 5.3.2: Training and Validation accuracy and Loss over the of Epochs (ViT)

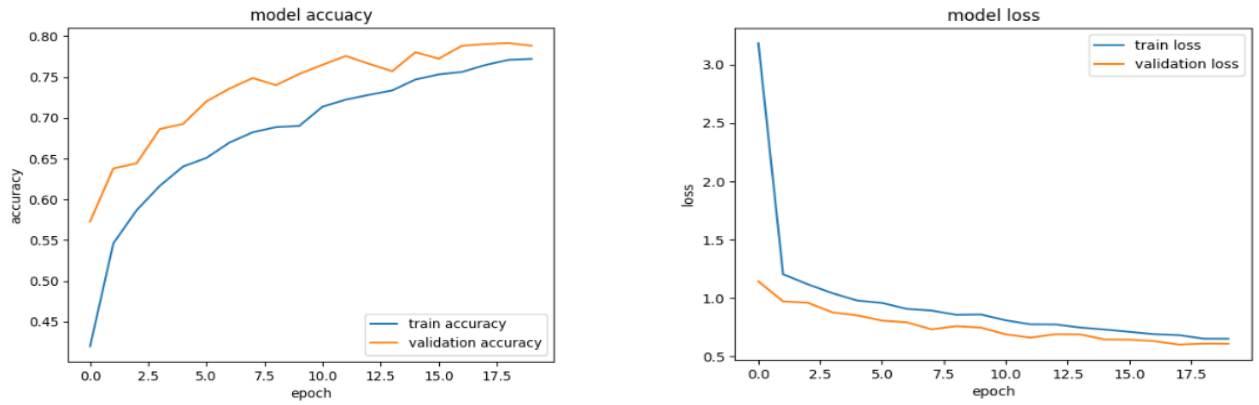


Figure 5.3.3: Training and Validation accuracy and Loss over the of Epochs (VGG19)

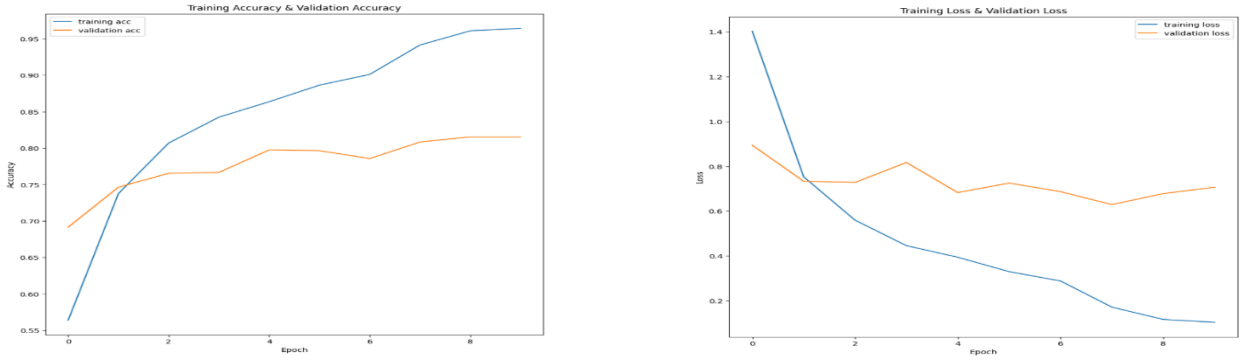


Figure 5.3.4: Training and Validation accuracy and Loss over the of Epochs (EfficientNet)

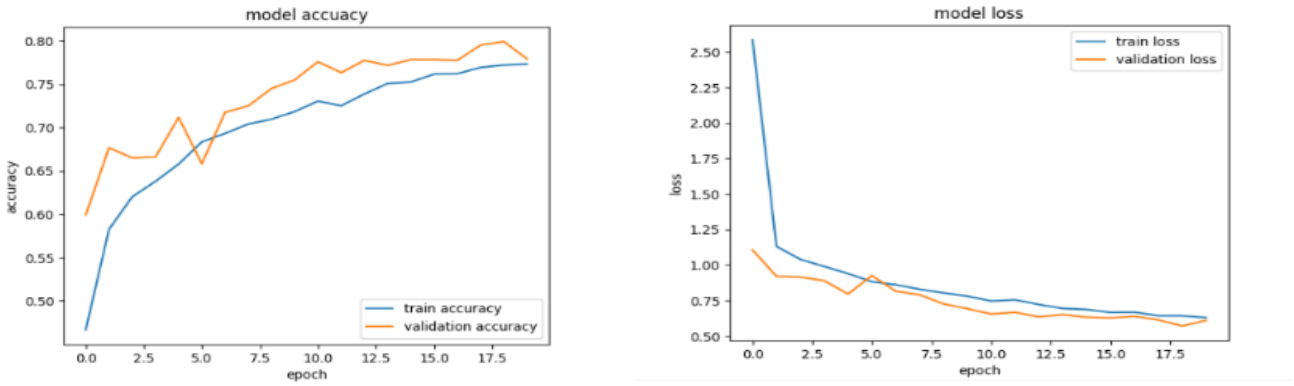


Figure 5.3.5: Training and Validation accuracy and Loss over the of Epochs (InceptionV3)

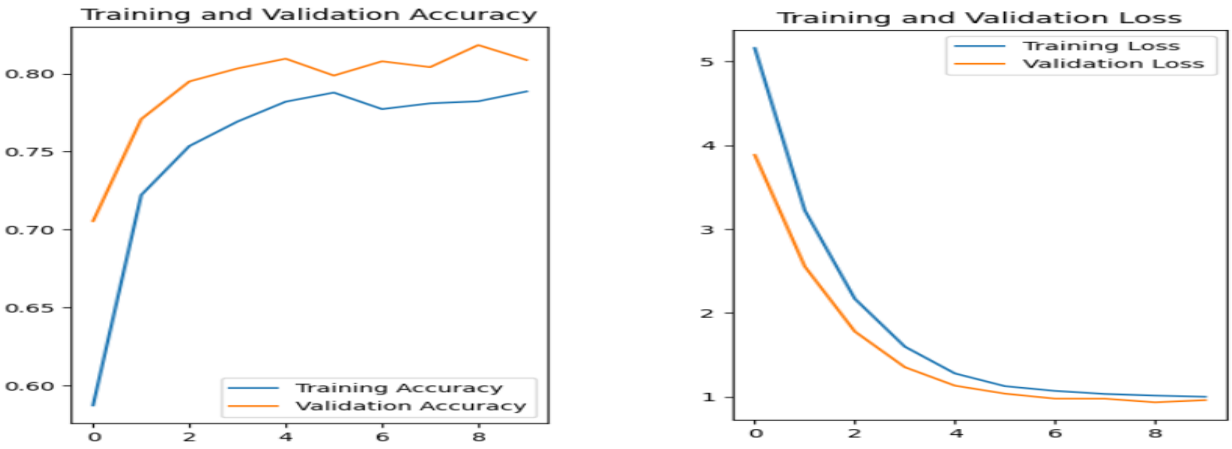


Figure 5.3.6: Training and Validation accuracy and Loss over the of Epochs (ResNet50)

The line chart shows the comparative performance of our proposed model (Vision Transformer) by depicting the accuracy and loss metrics for both the training and validation datasets across each epoch.

5.5 Summary

This extensive study evaluates the effectiveness of deep learning models in recognizing six fundamental emotions neutrality, anger, sorrow, pleasure, disgust, and surprise from face photographs. It underscores the crucial role of a large data set enriched with modern augmentation approaches as the research delves into Vision Transformer (ViT), ResNet50, VGG-19, InceptionV3, and Efficient-Net models.

ViT and Efficient-Net were the top performers, achieving accuracy scores of 82.96% and 82.36%, respectively. VGG-19 and InceptionV3 demonstrated accuracy of 79%, while ResNet50 obtained 81%, suggesting reasonable performance but with minor discrepancies in categorization. It indicates that transfer learning effectiveness falls when input photos deviate significantly from the training data in models like ImageNet. This is due to variations in facial expressions, background noise, and different augmentation procedures. It suggests that increasing the dataset size, especially with more diverse images, could improve the models' ability to predict emotions. Additionally, ensemble models—which combine different architectures show promise in achieving higher accuracy than individual models.

The results advocate for robust models like ViT and Efficient-Net and underscore the advantages of ensemble techniques and dataset extension in enhancing accuracy and reliability in emotion identification systems. In conclusion, ViT is the model we are proposing for FER.

CHAPTER 6

Impact on Society, Environment and Sustainability

6.1 Impact on Life

The development of precise emotion recognition has enormous potential for numerous areas of society, changing relationships between people and technology. We can develop more intuitive and compassionate interfaces by allowing robots to grasp human emotions more effectively. This innovation is incredibly disruptive in industries like healthcare. AI-powered systems may alter their answers depending on a patient's emotional state, thereby boosting the quality of patient care experiences. Imagine a situation where a virtual assistant can identify whether a patient feels uncomfortable during a telemedicine appointment, modify its tone, and reply appropriately, offering reassurance and assistance in real time. Similarly, emotion detection can tailor learning experiences by recognizing students' emotional reactions to instructional content. This helps instructors adjust their teaching approaches to fit individual learning styles and emotional requirements, boosting student involvement and comprehension. Furthermore, emotion recognition may change the user experience in entertainment by dynamically allowing games and interactive media to adapt to the user's emotional state. For example, a video game may modify its difficulty level according to the player's annoyance or boredom levels, delivering a more immersive and emotionally engaging experience.

6.2 Impact on Environment

While developing deep learning models for emotion detection provides vast potential, it also creates environmental constraints in part of the substantial computing resources necessary for training and deployment. Training complicated neural networks frequently includes large-scale data processing and considerable computational power, which may contribute to carbon emissions and energy consumption. This study addresses these environmental issues and presents an approach that seeks to find a compromise between obtaining high accuracy in emotion recognition and reducing the environmental footprint associated with model development and deployment. By investigating energy-efficient algorithms, improving hardware usage, and using cloud platforms given to renewable energy sources, we aim to reduce the environmental effect of our research efforts while advancing the area of emotion detection.

6.3 Ethical Aspects

In the search for technological progress, it is essential to keep ethical standards to ensure technology's responsible and fair use. This includes protecting user privacy and secrecy, getting educated permission to collect and use emotional data, and actively reducing bias in the development process. Central to our method is the adoption of solid data security, which means protecting user privacy and ensuring the secrecy of emotional data. We value openness and responsibility by clearly describing how emotional data will be used and kept, enabling users to make informed choices about their data. Furthermore, we employ data debiasing and fairness evaluation methods to ensure that our emotion recognition models are fair and equal across different groups, avoiding the continuation of biases and discrimination in automated decision-making.

6.4 Sustainability Plan

Our sustainability plan includes a diverse method to reduce the environmental effect of our study efforts while promoting long-term viability. We support open-source principles, sharing our code and models widely to promote teamwork and knowledge sharing across study groups, thereby lowering duplication of effort and improving resource usage. Additionally, we value sustainable practices in our research operations, including using energy-efficient computer tools and selecting cloud platforms committed to green energy sources. By adding sustainable principles into every part of our research, we hope to contribute to developing emotion recognition technology while reducing its environmental footprint and promoting responsible technological innovation.

6.5 Summary

This research explores the transformative potential of precise emotion recognition in healthcare, education, and entertainment. AI systems can enhance patient care by responding to emotional states, tailor learning experiences for better student engagement, and create more immersive entertainment by adapting content based on emotions. Addressing environmental concerns, the study proposes energy-efficient algorithms and renewable energy-powered cloud platforms to reduce the environmental impact of deep learning models. Ethical considerations focus on protecting user privacy, obtaining informed consent, and ensuring fairness in emotion recognition models. The sustainability plan includes promoting open-source collaboration, using energy-efficient computing tools, and selecting green energy cloud platforms to minimize the environmental footprint, advancing emotion recognition technology.

CHAPTER 7
Conclusion and Future Work

7.1 Conclusions

This research journey into emotion detection, integrating image processing and deep learning, has yielded significant insights and promising results. By harnessing these technologies, we've developed a tool poised to enhance the accessibility of emotion detection, enriching lives and advancing Human-Computer Interaction (HCI). Through meticulous evaluation of pre-trained deep learning models including Vision Transformer (ViT), VGG19, InceptionV3, Efficient-Net, and ResNet50, we've identified key human emotions, selecting Vision Transformer (ViT) for its remarkable accuracy of 82.93%.

This study marks a substantial leap in the application of advanced technology in HCI. The adoption of Vision Transformer (ViT) underscores its superior performance and potential impact on understanding human behavior through facial expressions. Looking ahead, these findings pave the way for integrating these techniques into practical applications, revolutionizing emotion detection and shaping a more resilient future in this field. The successful selection of Vision Transformer (ViT) represents a pivotal milestone, highlighting the potential of deep learning and image processing to redefine our understanding of emotions based solely on facial expressions.

7.2 Further Suggested Works

The implications for further study stemming from this research are wide-ranging and offer numerous avenues for deeper exploration and advancement in the field of emotion detection and human-computer interaction (HCI). Firstly, future research endeavors could focus on refining and optimizing the selected deep learning model Vision Transformer (ViT) to achieve even greater accuracy and efficiency in emotion detection. This might involve fine-tuning model parameters, exploring alternative architectures, or integrating additional data sources to enhance the model's performance. Secondly, expanding and diversifying the dataset used for training and testing emotion detection models could lead to more robust and generalizable results. Collecting data from a broader range of demographics, cultural backgrounds, and emotional expressions could help mitigate biases and improve model applicability across diverse populations.

Additionally, exploring the integration of multiple components in addition to facial expressions like voice tone, and body-language could enhance the accuracy and richness

of emotion detection systems. Studying the ethical implications of emotion detection technology and HCI applications is also crucial for responsible development and deployment. Further research could delve into issues such as privacy concerns, data security, algorithmic biases, and the potential impact on human autonomy and well-being. Overall, continued research in these areas has the potential to advance the state-of-the-art in emotion detection and HCI, fostering more accurate, adaptive, and ethically responsible human-computer interactions in various contexts.

7.3 Limitations/ Conflict of Interests

This study on emotion recognition using deep learning has several limitations and potential conflicts of interest. The models' performance heavily depends on the quality and diversity of the training data, which may not be representative of diverse populations, leading to biases. These models might not generalize well to real-world scenarios with varying conditions. Ethical concerns include privacy and consent in facial data collection. High computational resources are required, limiting accessibility and scalability, especially in resource-constrained environments. Real-time implementation remains challenging due to the need for optimization for low latency and efficient processing on edge devices. The "black box" nature of deep learning models makes decision interpretation difficult, which is a significant limitation in critical applications.

Researcher’s affiliations with commercial entities may introduce biases, necessitating transparency and disclosure. Variability in reported performance metrics across studies highlights the need for standardized metrics and benchmarking for fair comparisons and reproducibility. Addressing these limitations requires improving dataset diversity, developing more interpretable models, and adhering to ethical guidelines in facial data use. Collaborative efforts and technological advancements are essential to enhance the reliability and applicability of emotion recognition systems.

References:

[1] Gautam, C., & Seeja, K. R. (2023). Facial emotion recognition using Handcrafted features and CNN. Procedia Computer Science, 218, 1295-1303.

[2] Avanija, J., Madhavi, K. R., Sunitha, G., Sangapu, S. C., & Raju, S. (2022, March). Facial Expression Recognition using Convolutional Neural Network. In 2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR) (pp. 17). IEEE.

[3] Kanagaraju, P., Ranjith, M. A., & Vijayasarathy, K. (2022). Emotion detection from facial expression using image processing. International Journal of Health Sciences, 6(S6), 1368–1379.

[4] Dachapally, P. R. (2017). Facial emotion detection using convolutional neural networks and representational autoencoder units. arXiv preprint arXiv:1706.01509.

[5] Syed, W. A., Uppalapti, A. K., Vadakattu, A., Vasireddy, A., & Reddy, P. C. (2018). Emotion Detection through Facial Feature Recognition. Emotion, 5(02).

[6] Ko, B. C. (2018). A brief review of facial emotion recognition based on visual information. sensors, 18(2), 401.

[7] Jaiswal, A., Raju, A. K., & Deb, S. (2020, June). Facial emotion detection using deep learning. In 2020 international conference for emerging technology (INCET) (pp. 1-5). IEEE.

[8] Ali, M. F., Khatun, M., & Turzo, N. A. (2020). Facial emotion detection using neural network. the international journal of scientific and engineering research.

[9] Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). SN Applied Sciences, 2(3), 446.

[10] Saxena, A., Khanna, A., & Gupta, D. (2020). Emotion recognition and detection methods: A comprehensive survey. Journal of Artificial Intelligence and Systems, 2(1), 53-79.

[11] Modi, S., & Bohara, M. H. (2021, May). Facial emotion recognition using convolution neural network. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1339-1344). IEEE.

[12] Gajarla, V., & Gupta, A. (2015). Emotion detection and sentiment analysis of images. Georgia Institute of Technology, 1, 1-4.

[13] Alonso-Martin, F., Malfaz, M., Sequeira, J., Gorostiza, J. F., & Salichs, M. A. (2013). A multimodal emotion detection system during human–robot interaction. Sensors, 13(11), 15549-15581.

[14] Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016, March). Going deeper in facial expression recognition using deep neural networks. In 2016 IEEE Winter conference on applications of computer vision (WACV) (pp. 1-10). IEEE.

[15] Madinah, S. A. (2017). Emotion detection through facial feature recognition. International Journal of Multimedia and Ubiquitous Engineering, 12(11), 21-30.

[16] Mamieva, D., Abdusalomov, A. B., Kutlimuratov, A., Muminov, B., & Whangbo, T. K. (2023). Multimodal Emotion Detection via Attention-Based Fusion of Extracted Facial and Speech Features. Sensors, 23(12), 5475.

[17] Turabzadeh, S., Meng, H., Swash, R. M., Pleva, M., & Juhar, J. (2018). Facial expression emotion detection for real-time embedded systems. Technologies, 6(1), 17.

[18] Uddin, M. Z., Khaksar, W., & Torresen, J. (2017). Facial expression recognition using salient features and convolutional neural network. IEEE Access, 5, 26146-26161.

[19] Joseph, A., & Geetha, P. (2020). Facial emotion detection using modified eyemap–mouthmap algorithm on an enhanced image and classification with tensorflow. The Visual Computer, 36(3), 529-539.

[20] Alreshidi, A., & Ullah, M. (2020, February). Facial emotion recognition using hybrid features. In

Informatics (Vol. 7, No. 1, p. 6). MDPI.
[21] The Vision Transformer model, MachineLearningMastery.com, available at <<https://machinelearningmastery.com/the-vision-transformer-model/>>, last accessed on 23-06-2024 at 12:00 PM.
[22] VGG-Net architecture explained, Medium, available at <<https://medium.com/@siddheshb008/vgg-net-architecture-explained-71179310050f>>, last accessed on 23-06-2024 at 12:15 PM.
[23] Inception V3 model architecture, OpenGenus IQ: Learn Algorithms, DL, System Design, available at <<https://iq.opengenus.org/inception-v3-model-architecture/>>, last accessed on 23-06-2024 at 12:30 PM.
[24] Exploring Resnet50: An in-depth look at the model architecture and code implementation, Medium, available at <<https://medium.com/@nitishkundu1993/exploring-resnet50-an-in-depth-look-at-the-model-architecture-and-code-implementation-d8d8fa67e46f>>, last accessed on 23-06-2024 at 12:45 PM.
[25] EfficientNet and its performance comparison with other Transfer Learning Networks, Wisdom ML, available at <<https://wisdomml.in/efficientnet-and-its-performance-comparison-with-other-transfer-learning-networks/>>, last accessed on 23-06-2024 at 1:00 PM.
[26] Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. Remote Sensing, 13(3), 516.
[27] Sudha, V., & Ganeshbabu, T. R. (2021). A Convolutional Neural Network Classifier VGG-19 Architecture for Lesion Detection and Grading in Diabetic Retinopathy Based on Deep Learning. Computers, Materials & Continua, 66(1).
[28] Burkapalli, V. C., & Patil, P. C. (2020). TRANSFER LEARNING: INCEPTION-V3 BASED CUSTOM CLASSIFICATION APPROACH FOR FOOD IMAGES. ICTACT Journal on Image & Video Processing, 11(1).
[29] Liang, J. (2020, September). Image classification based on RESNET. In Journal of Physics: Conference Series (Vol. 1634, No. 1, p. 012110). IOP Publishing.
[30] Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR.

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title:
Emotion That Speaks: Peering Beyond the Obvious with Deep Learning for Emotion Recognition

Student ID: 203-15-14472, 203-15-14486

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge on emotion detection from facial expression for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish these goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**

CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**
CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project demonstrates **fundamental engineering (K3)** principles by employing deep neural networks, data augmentation techniques, and diverse CNN architectures for image processing and classification tasks. The project demonstrates **specialist knowledge (K4)** by conducting transfer learning with advanced CNN architectures like VGG19, ResNet, InceptionV3, and Efficient-Net and also Transformer architecture ViT enhancing emotion detection accuracy from facial expression by using images.	**Page no:** [15-25] **Section:** [3.2, 3.3] **Page no:** [19-25] **Section:** [3.3]

				The project applies **engineering practice & design (K5)** by the figure of process of experiments. The project addresses **engineering practice & technology (K6)** by employing transfer learning models and transformer model.	**Page no:** [35, 39] **Section:** [5.3, 5.5] **Page no:** [19-25] **Section:** [3.3]
				This project ensures to **K8 (Research Literature)** by synthesizing insights from recent studies, to advance emotion detection using deep learning, showcasing a comprehensive understanding of current methodologies.	**Page no:** [8-11] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting Requirements**	Yes	CO2, and CO7	This project addresses **EP-2** by recognizing the hurdles in emotion detection, including limitations of traditional methods and the complexities of integrating transfer learning and Vision Transformer model. Through comparative analysis, it confronts challenges in understanding spatial distributions, offering insights for refining diagnostic methodologies.	**Page no:** [3-4] **Section:** [1.5] **Page no:** [35, 39] **Section:** [5.3, 5.5]

3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This project addresses **EP-3** by meticulously comparing experimental outcomes, highlighting Vision Transformer model as the chosen significant solution for enhancing emotion detection amidst multiple potential approaches.	**Page no:** [32-38] **Section:** [5.2, 5.3]
4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's interdisciplinary approach extends beyond computer science and engineering, impacting human interactions and understanding of human emotions, contributing to advancements in HCI practices which indicates **EP-4**.	**Page no:** [3] **Section:** [1.4]
5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders involved and conflicting requirements**	Yes	CO8	As part of our project, we have to visit and a psychologist. The purpose is to understand how we can determine emotion based on a person's expression. This meeting will guide us in determining which facial expressions convey what type of emotion. This indicates **EP-6**	**Page no:** [13] **Section:** [3.1]

7.	**EP7: Interdependence**	Yes	CO5	This project's comprehensive approach addresses high-level problems by integrating various components across data collection, statistical analysis, and proposed methodology, ensuring a holistic solution to complex challenges in emotion detection which ensures **EP-7**.	**Page no:** [13-25] **Section:** [3.1, 3.2, 3.3]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, deep learning frameworks to ensure systematic research and contribute to advancements in emotion detection through transfer learning and Vision Transformer model	**Page no:** [25-26] **Section:** [3.4]
2.	**EA2: Level of interaction**	No		N/A	N/A

3.	**EA3: Innovation**	No		N/A	N/A
4.	**EA4: Consequences for society and the environment**	Yes		This project contributes to society by improving HCI through advanced emotion detection methods, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for patient data privacy.	**Page no:** [40-41] **Section:** [6.1, 6.2, 6.3, 6.4]
5.	**EA-5: Familiarity**	Yes		This project expands upon existing research by examining a novel approach in emotion detection through transfer learning and Vision Transformer model demonstrated through preliminary terminologies and a comprehensive comparative analysis, offering new insights into the field.	**Page no:** [6, 10-12] **Section:** [2.1, 2.3, 2.4, 2.5]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle.	**Page no:** [26-28] **Section:** [3.5]

2	CO9	Yes	The project demonstrates adherence to ethical principles by prioritizing the participant's privacy, obtaining informed consent, and transparently documenting the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced emotion detection technologies which comply with **CO9**.	**Page no:** [41] **Section:** [6.3]
3	CO10	No	N/A	N/A
4	CO12	Yes	The project's dedication to continuous learning **(CO12)** and adaptation within the dynamic technological landscape is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges.	**Page no:** [15-25, 32-38, 42-43] **Section:** [3.2, 3.3, 5.2, 5.3, 7.2]

PLAGIARISM REPORT

Emotion That Speaks: Peering Beyond the Obvious with Deep Learning for Emotion Recognition.

ORIGINALITY REPORT

15%	9%	7%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Submitted to Higher Education Commission Pakistan Student Paper	2%
3	Submitted to Daffodil International University Student Paper	2%
4	mdpi-res.com Internet Source	1%
5	www.ncbi.nlm.nih.gov Internet Source	<1%
6	"Advanced Computing and Intelligent Technologies", Springer Science and Business Media LLC, 2024 Publication	<1%
7	Submitted to University of Hertfordshire Student Paper	<1%
8	www.mdpi.com Internet Source	<1%