Recognition of Affective Emotional States using Facial Expressions

by

Kazi Shashwata Rahman
20101167
Annesha Karim
20101048
Mostafa Al Mamun Fahim
20101366

A thesis report submitted to the Department of Computer Science and Engineering in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
May 2024

Declaration

It is hereby declared that

1. The thesis report submitted is our own original work while completing degree at Brac University.

2. The thesis report does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis report does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Kazi Shashwata Rahman
20101167

Annesha Karim
20101048

Mostafa Al Mamun Fahim
20101366

Approval

The thesis report titled "Recognition of Affective Emotional States using Facial Expressions" submitted by

1. Kazi Shashwata Rahman (20101167)
2. Annesha Karim (20101048)
3. Mostafa Al Mamun Fahim (20101366)

Of Spring, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on May 1st, 2024.

Examining Committee:

Supervisor:

Beig Rajibul Hasan
Lecturer
Department of Computer Science and Engineering
BRAC University

Co-Supervisor:

MD Mehedi Hasan Shawon
Lecturer
Department of Electrical and Electronic Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam, PhD
Associate Professor
School of Data and Sciences
BRAC University

Head of the Department:
(chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
School of Data and Sciences
Brac University

Ethics Statement

We therefore affirm that this thesis is founded upon the results obtained from our investigation. In the text, all additional sources of information have been duly acknowledged. No previous submission of this thesis, in its entirety or in part, for the conferral of a degree to any other university or institute has been documented.

Abstract

Affective computing is becoming more important, and one of the things that is becoming more important is the analysis of face expressions. This is one of the area issues that is getting a lot of attention. Using this way of analysis, you can figure out how someone is feeling. Human-computer connection, social machines, and mental health care, to name a few, are all affected. This study shows how convolutional neural networks (CNNs), which are a type of artificial neural network, can be used to find out about affective emotional states from face expressions. CNNs can utilize convolutional neural network features. A good example of an artificial neural network is the news network CNN. The suggested system already uses very effective preprocessing methods that can deal with a wide range of problems, such as differences in lighting, posture, and occlusion. The system already has these tools built in. These ways of working are already built into the system. Because of these methods, it is now possible to do statistical analyses that can be trusted more. This makes it possible to compare facial images in ways that weren't possible before. CNN's design is great at capturing the complicated spatial patterns and relationships that are part of facial expressions. This is just one of the many things the network can do. This is because CNN was made so that it could do all of these things. So, it's easy to spot things that are different from a lot of other things. Extensive experimental reviews on world-famous benchmark datasets show that, when it comes to picking up on subtle emotional signals, the system is better than state-of-the-art methods in terms of both accuracy and precision. The machine was in charge of making these assessments. Some of the most knowledgeable experts on artificial intelligence in the world, from many different countries, did these reviews. Because the suggested framework is flexible and easy to use in a wide range of situations, it is best for use in the real world. Because of this, it has the ability to be very useful in a wide range of industries. These tools could change what happened in the past. In the future, researchers may look into multimodal methods that include more modalities, such as sound and body movements, build a framework for real-time implementation, and add more emotional states to the dataset. The framework could also be changed to make deployment easier in real time. This could be done as part of a multimodal approach, which uses different ways to talk to people. This is possible because of multimodal methods. This work opens the door for future progress in many areas, such as human-computer interaction, social robotics, and mental health support systems, by showing how well CNNs can predict affective emotional states from facial expressions. This is done by putting the spotlight on how well CNN is at predicting how people will feel. This shows how well CNNs can figure out how people feel just by looking at their faces. Deep learning is used in this method to make it easier to create applications that can recognize and react to a wide range of human emotions. People can get help and support from these groups in a number of different situations.

Keywords: facial expression, emotion recognition, CNN, RESNET, INCEPTIONRESNET, VGGNET

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The proliferation and swift advancement of computer systems, software, and networks have had a profound impact on our everyday existence, enhancing the ease and efficacy of various tasks. Facial emotion recognition systems have become increasingly significant in contemporary times due to their ability to capture human behavior, intentions, and emotions. Facial emotion recognition systems employ computer vision and artificial intelligence (AI) algorithms to investigate and interpret real-time facial emotions. The algorithms utilized in this study are capable of accurately identifying and classifying a range of emotions, such as happiness, sadness, anger, surprise, and others. This is achieved through the analysis of key facial features, including eyebrow movements, eye dilation, lip shape, and overall muscular activity. The technology in question holds great potential in various industries such as healthcare, entertainment, marketing, and security.

The implementation of facial emotion recognition systems has the potential to aid healthcare professionals in the diagnosis and monitoring of various mental health conditions such as anxiety, sadness, and autism spectrum disorders. The utilization of these technologies has the potential to yield valuable insights into the emotional states of patients through precise capture and analysis of facial expressions. This, in turn, may facilitate the customization of treatment plans and interventions by healthcare practitioners. The utilization of facial expression recognition technology could potentially aid in telemedicine scenarios by providing insight into patients' emotional states during virtual consultations, where physical cues may be limited.

Facial expression detection technologies have been utilized by the entertainment sector to enhance user experiences. The aforementioned technology is utilized to generate verisimilar and immersive environments within video games and virtual reality software. It has been suggested that virtual avatars and gaming characters have the potential to respond in real-time to players' emotions and facial expressions, thereby modifying their behavior and conversation to offer a more personalized and engaging experience. Face expression recognition technology can be utilized by filmmakers and marketers to identify audience preferences and reactions. This information can then be used to adjust their content and advertising strategies in order to elicit the desired emotional responses.

The implementation of facial emotion recognition systems in marketing is crucial for obtaining valuable insights into the preferences and behaviors of customers. Real-time feedback on the emotional impact of products may be obtained by retailers through the observation of consumers' facial expressions during product use. By leveraging this data to inform product ideation, marketing strategies, and enhancements to the customer journey, promotional initiatives can achieve greater efficacy and precision.

Face emotion detection technologies are extensively utilized in security and surveillance applications. Through the process of facial expression analysis, researchers can potentially identify individuals who may pose a threat or exhibit suspicious behavior in public settings. This method involves observing and interpreting facial cues that may indicate feelings of hostility, fear, or distress. The implementation of this technology has the potential to enhance security measures and mitigate potential security breaches in high-security areas such as airports and railway stations.

It is imperative to exercise responsible usage of facial expression detection technologies and to take into account any potential ethical concerns. When utilizing this particular technology, it is imperative to uphold the principles of privacy, obtain explicit consent from individuals, and implement robust data protection protocols.

The speed and accuracy of conventional methods for facial emotion recognition are often limited. The utilization of deep learning methodologies, specifically via deep convolutional neural networks (CNNs), has demonstrated a more effective strategy. The objective of this study is to construct a convolutional neural network (CNN) with significant depth that can effectively identify five distinct emotional expressions exhibited by human faces. The model in question exhibits promise for various applications, such as customer feedback analysis and face unlocking, where accurate interpretation of facial expressions is of utmost importance. The predominant mode of human communication is oral language, however, nonverbal cues such as body language and facial expressions are also integral components that serve to reinforce speech and express emotions. Non-verbal communication is a crucial aspect of interpersonal relationships, with facial expressions playing a significant role in this domain. The potential applications of automatic recognition of facial expressions are vast and varied, spanning multiple fields such as natural human-machine interfaces, behavioral science, and clinical practice. The recognition of facial expressions is a task that humans perform with ease, but the development of machines capable of reliably recognizing expressions remains a challenge. The present study suggests a methodology that relies on convolutional neural networks (CNNs) to recognize facial expressions. The employed system utilizes convolutional neural networks (CNNs) to predict the corresponding facial expression label, such as anger, happiness, fear, sadness, disgust, or neutral, based on an input image. Prior studies have made notable progress in the field of automatic expression classification, employing methodologies such as face detection, feature extraction, and diverse classification techniques. This study aims to investigate the efficacy of convolutional neural networks (CNNs) in accurately recognizing facial expressions, utilizing the latest technological advancements in the field. The approach based on convolutional neural networks (CNNs) that was presented exhibits potential for enhancing the automation of the task at hand, and adds to the continuous advancement of reliable systems for recognizing facial expressions.

We enhanced the VGGNET model to precisely cater to the task of emotion classification by training it using our own dataset. The model achieved an accuracy of 87.53%. To enhance the training process, we implemented various data augmentations on the training data. These augmentations encompassed modifications to the height and breadth, rotation, shear, channel shift, and horizontal flip. Furthermore, we integrated 6 customised layers into the foundational VGGNET model. The careful and precise technique, along with the high accuracy achieved, demonstrates the effectiveness of our methodology in recognising facial emotions. The model's impressive performance can be ascribed to the utilisation of a meticulously curated dataset and the incorporation of data augmentations. The addition of personalised layers to the basic model has greatly enhanced its ability to understand the complexity of human emotions. In conclusion, our research represents a significant progress in the field of facial emotion recognition, showcasing the efficacy of combining meticulous dataset preparation with advanced machine learning techniques. The results of our study offer promising prospects for additional research and real-world application in this domain.

In light of recent advancements and widespread adoption of computer systems, software, and networks, facial expression detection systems have emerged as a powerful means of capturing human behavior, intentions, and emotions. The technology exhibits a wide range of potential applications across various domains including healthcare, entertainment, marketing, and security. The technology in question holds significant potential for improving our standard of living and promoting a deeper comprehension of human emotions and connections.

1.2 Problem Statement

The present study centers on the constraints of traditional approaches employed for the identification of facial emotions, particularly their restricted precision and velocity. The growing importance of computer systems and networks in our daily routines has made it imperative to accurately capture human emotions through facial expression recognition. According to the researchers [6] [12], the utilization of deep learning techniques, specifically deep convolutional neural networks (CNNs), presents a more advanced strategy for facial emotion recognition in contrast to conventional methods. The aim of this study is to create a custom convolutional neural network (CNN) model with a high level of accuracy in identifying and categorizing seven distinct human facial emotions. The model under consideration will undergo training using a dataset of facial images that have been manually collected from the tv show "Friends". These images were captured from videos and manually organized. The principal objective of this study is to improve the precision and effectiveness of facial emotion recognition systems, which could have practical implications in the fields of customer feedback analysis. The present study aims to tackle the issue of dependable and precise machine-based recognition of facial expressions. To this end, a custom Convolutional Neural Network (CNNs) is proposed as a potential solution.

The aim of this study is to create a system for recognizing facial expressions using a convolutional neural network (CNN) and to assess its efficacy. The input to the system is an image, and the system predicts the corresponding label for the facial expression. The predicted label encompasses a range of emotions, including anger, happiness, fear, sadness, disgust, surprise and neutral. The present study presents a comprehensive survey of the existing literature on expression recognition, encompassing various approaches such as classifiers based on prototypical emotions and individual muscle movements. In addition, the CNN architecture is expounded upon, which includes convolutional layers, fully connected layers, and various techniques such as dropout, and max-pooling.The primary aim of this study is to evaluate the efficacy of the proposed system and make a significant contribution to the progress of facial expression recognition methodologies. The present research study focuses on the difficulties associated with recognizing facial emotions and introduces innovative techniques that employ deep learning and convolutional neural networks (CNNs). The objectives of the research involve the creation of precise and effective models, assessment of their performance, and investigation of their potential applications across diverse fields.

1.3 Research Objective

The present study aims to overcome the constraints of traditional facial emotion recognition approaches by highlighting the benefits of employing deep learning methodologies, specifically deep convolutional neural networks (CNNs). The primary aim of this study is to construct a resilient and efficient convolutional neural network (CNN) architecture that can precisely identify and categorize various human facial emotions. Through the utilization of deep learning algorithms and the implementation of a manually curated dataset of facial images, the objective is to substantially improve the efficiency and precision of facial emotion recognition systems.

The study recognizes the increasing significance of facial emotion recognition in the current technology-oriented era, given that these systems possess the capability to acquire and evaluate valuable data regarding human behavior, emotions, and motives. The proposed system seeks to address the shortcomings of conventional methods by utilizing deep learning techniques. These conventional methods are often plagued by slower processing times and reduce the accuracy levels. The aim of this study is to showcase the effectiveness of utilizing deep learning techniques, particularly a deep CNN model, for the purpose of facial emotion recognition. The study seeks to establish that this approach provides superior accuracy and efficiency compared to other methods.

This study makes a valuable contribution to the field of facial expression recognition through its examination of multiple convolutional neural network architectures and its exploration of techniques such as dropout and max-pooling. The aim of this study is to enhance the performance of the proposed system by analyzing various network architectures and feature extraction mechanisms.

Furthermore, the study aims to investigate the pragmatic implementations of the newly created facial emotion recognition technology. The identified use cases, namely customer feedback analysis and face unlocking, underscore the significance and applicability of the research. The precise identification of facial emotions has the potential to provide businesses with valuable insights into customer interactions. Additionally, it can facilitate the development of secure and convenient face unlocking mechanisms, thereby enhancing user experiences.

This study aims to create and test deep learning-based facial emotion identification algorithms. The study aims to improve recognition system accuracy and efficiency. The study examines how said systems are used in consumer feedback analysis, face unlocking, natural human-machine interactions, behavioural science, and therapeutic practice.

Chapter 2

Literature Review

Zhao and Pietikainen (2007) [1] studied face expression identification using the LBP-TP approach. Scientists call visual patterns with temporal fluctuations "dynamic textures" like undulating waves or meandering smoke. The authors recommend using Local Binary Patterns with Three Orthogonal Planes (LBP-TP) as a feature descriptor to capture spatiotemporal information in dynamic textures. The LBP-TP method generates binary codes by evaluating neighboring pixel brightness levels over many planes. The above scripts generate histograms showing the frequency distribution of discrete texture patterns throughout dynamic sequences. Facial expression recognition tests LBP-TP's dynamic texture recognition. The algorithms' performance is assessed using published datasets and compared to other advanced methods. The summary lacks quantitative data, however the paper should describe the precision. The research found that LBP-TP can accurately recognize face expressions and capture their underlying dynamic patterns. The proposed approach has potential in various sectors, such as human-computer interaction, emotion recognition, and monitoring.

Zhao et al. (2011) studied the effectiveness of NIR movies in facial expression recognition [2]. In low light, NIR imaging is tried for face expression recording. Scholars extracted, selected, and categorized features. These strategies offer precise recognition when combined. The approach's accuracy, precision, recall, and F1-score were assessed using the CK+ dataset. The study showed that NIR imaging can accurately detect face expressions in low light or obstructions. The study showed that NIR recordings detect face muscle activity better than visible-light imagery. Local binary and ternary patterns (LBP and LTP) resist changing lighting conditions for feature extraction. Adaboost feature selection improves discrimination. Classification using SVMs. This study examines how NIR movies can detect facial expressions. This study shows that the proposed architecture and methods can classify face expressions. This research may help HCI and emotional computing. It aims to create facial expression recognition. Investigating low illumination.

Schroff et al.[4] suggest an approach that combines face verification, recognition, and clustering by extracting representations from raw pixel data. Two deep convolutional network topologies are studied for their internal operations. The initial architecture, influenced by Zeiler & Fergus, uses many convolutional layers, nonlinear activation functions, local response normalization, and maximum pooling to achieve the desired output. Prior studies and one-dimensional convolutional layers are used. Inception is a great computer vision design. Work combines convolutional and pooling layer results using mixed layers. The results show that the models could perform similarly with less parameters and computations. The research presents a pixel-based method for validating, recognizing, and categorizing facial photographs. The study found that the fixed center crop method provides a remarkable classification accuracy of 98.87%. Furthermore, the study shows a record-high mean standard error of 99.63% when using the additional face alignment technique.

Alizadeh et al. (2017) investigated facial expression recognition using CNNs and deep learning algorithms [6]. Authors stress accurate photo and video face expression categorization. Traditional face expression recognition uses traits. Local Binary Patterns with hand-built Gabor filters. Complex spatial features and setups challenge these approaches. CNNs help scientists overcome restrictions. CNN hierarchizes data automatically. Machine learning restrictions are addressed in this paper. CNNs extract facial features using convolutional, pooling, and fully connected layers. The authors closely examine these strata. Layers' roles in this process are investigated. Microexpression databases are small, thus scientists sought strategies to enlarge them. CNN model durability and generalization are examined here. Rotation, horizontal flipping, and random cropping are tested. The study compares CNNs with deep learning for facial emotion recognition. CNNs capture facial expressions better due of their intricacy. Studies how random cropping, horizontal flipping, and rotation improve CNN model generalization and stability.

In 2017, Tarnowski et al. [7] examined the efficacy of facial expressions in recognizing emotions and the accuracy of categorization algorithms. This study contrasts subject-dependent vs subject-independent classification. The former trains and tests models on the same people, the latter on different groups. The MLP classifier has 90% subject-dependency accuracy, while the 3-NN has 96%. Both algorithms classify emotions from facial expressions on identical persons. Subject-independent MLP classifier accuracy is 75.9%. Interesting, the 3-NN classifier outperforms it by 95.5%. Multi-person emotion detection is good with 3-NN. Comparing "natural" subject-based and "random" partitioning helps us understand how data partitioning affects model accuracy. The results reveal that "random" division is more accurate than "natural" division. Subject-dependent classification using "natural" division yields 73% accuracy for the MLP classifier and 70% for the 3-NN classifier. For subject-independent classification, MLP classifier achieves 73% "natural" division accuracy. In contrast, the 3-NN classifier is 63% accurate. The study found facial expression-based emotion classification reliable. MLP and 3-NN classifiers excelled independently and subject-dependently. To enhance emotion recognition algorithms, examine data partitioning methods.
Chen et al. (2019) [8] explored the growing importance of facial emotion recognition in computer vision research. Social and emotional expression are examined. Older manual feature extraction methods used Gabor Wavelets and sparse coding. Neutral face removal improves recognition in several ways. CNNs changed this field. These networks became skilled through deep learning. Research used global appearance

data, local geometry cues, facial landmarks, and optical flow to improve recognition. Facial region localization was examined to improve feature extraction. Work began to comprehend Action Unit mechanics. Although deep learning algorithms have improved, problems remain. The study found domain knowledge, local geometry, and GAN-based approach limitations. The study developed unique Facial Motion Prior Networks. FMPN shows expressive face motion with the Face-Motion Mask Generator. Previous Fusion Net uses historical data. FMPN tops CK+, MMI, AffectNet. Moving muscle mask guidance and pre-calculated ground-truth masks help. This study suggests domain expertise and technique limits improve face expression recognition.

Ilham et al. [23] investigated the integration of AI and psychology in Facial Expression Recognition (FER). Deep learning may give smart psychiatry therapies. Deep learning enables clinical identification and therapy customisation. AFEA and FEA are compared by psychiatrists. A quantitative system must automatically assess facial expressions. AFEA numerically analyzes facial expressions. Modern approaches pass AFEA. Deep learning uses layers to study hierarchies. Light, posture, and face emotions have not modified layer properties. Neural networks value usability. Deep neural layers withstand disturbance. Faces are understood via deep learning. They sense emotions, intentions, and conduct. Deep convolutional neural networks recognise faces. These systems employ facial recognition to measure user response and unlock goals. Deep learning mimics brain function via neural networks. Two-dimensional CNNs correctly recognize five facial emotions in research subjects. System training dataset was manually collected using mobile phone camera.

Recent computer vision research focuses on facial emotion identification. Academic study by Sultana.A et al (2023) [26] may have influenced the observed development. As stated in the introduction, facial expressions are essential to communication. Natural language processing, psychiatry, gaming, and marketing have examined it. Facial expression problems for emotion deduction are examined in this study. The research was restricted by limited datasets, unpredictable environmental variables, insufficient training data for various emotional states, and lighting and posture. TensorFlow was used to study face expressions by Sultan et al. (2023). The objective was to understand emotions. The study covers image collecting and preprocessing, the VGG19-based deep convolutional neural network (DCNN), and iterative training and testing to fine-tune the network's parameters. The VGG19 model extracts image features using convolutional and fully-connected layers. The suggested model had 94.8% and 93.7% accuracy on the Extended Cohn-Kanade (CK+) and Japanese Female Facial Expression (JAFFE) datasets, respectively. The approach was extensively tested on a big dataset. It could be utilized in e-learning, surveillance, and healthcare infrastructure. Application may improve these areas.
Facial emotion recognition (FER) plays a crucial role in enhancing communication between computers and humans. Convolutional Neural Networks (CNNs) are a type of deep learning method that can automatically extract features and are highly efficient in terms of computation[15]. They show great potential for this particular task. By utilizing the VGGNet architecture as a standalone network, the research was able to reach a remarkable accuracy of 73.28% on the FER2013 dataset, without the need for additional training data. This accuracy is considered to be the highest

in the industry.

This study [18] builds upon previous research that highlights the importance of hyperparameter tuning in improving model performance. The specific goal of this study is to optimize the hyperparameters of Convolutional Neural Networks (CNNs) for facial emotion recognition. The research emphasizes the importance of hyperparameter optimization strategies, such as Random Search, in improving the accuracy of deep learning models. This study follows the same paradigm.

The difficulty of face emotion recognition stems from the intrinsic variety in facial architecture and the ambiguous nature of emotions[17]. Previous studies utilizing Convolutional Neural Networks (CNNs) for emotion recognition have shown insufficient effectiveness in comparison to tasks such as object detection. The research suggests a multi-task learning strategy that surpasses current cutting-edge algorithms by including emotion detection, gender, age, and ethnicity into a single CNN model.

The article "Facial Emotion Recognition using Convolutional Neural Networks" by Akash Sarava et al. [11] introduces a model based on Convolutional Neural Networks (CNN) to identify seven basic emotions based on facial expressions. The suggested model achieves a final accuracy of 0.60 and is composed of six convolutional layers, two max pooling layers, and two fully connected layers. The research emphasizes the benefits that Convolutional Neural Networks (CNNs) provide in terms of capturing the spatial characteristics of facial expressions. CNNs demonstrated superior performance compared to traditional neural networks and decision trees in this particular task. The research greatly advances the field of facial emotion identification by building a strong Convolutional Neural Network (CNN) model that is suitable for further examination.

Authored by H.M. Shahzad et al [25], the study "Hybrid Facial Emotion Recognition Using CNN-Based Features" describes a technique for recognizing facial emotions by combining convolutional neural networks (CNNs) and machine learning (ML) methods. The researchers extracted features from convolutional neural networks (CNNs) that had already been trained and evaluated several machine learning (ML) algorithms in order to accomplish classification tasks. The research findings revealed that the implementation of ensemble and Support Vector Machine (SVM) classifiers instead of the traditional SoftMax classifier resulted in improved emotion recognition accuracy across the FC6, FC7, and FC8 layers of Deep Convolutional Neural Networks. The results of the study suggested that when Convolutional Neural Networks (CNNs) are combined with other Machine Learning (ML) techniques, they can produce more effective results when applied to emotion perception tasks. The research also underscores the potential for improving the accuracy of image classification through the utilization of alternative classifiers and pre-trained convolutional neural networks (CNNs).

The paper "Emotion Recognition Using Convolutional Neural Networks" by Chieh-En James Li [9] et al presents a detailed neural network model designed to accurately identify facial emotions . The authors get the most advanced findings by enhancing the performance of a pre-trained VGG-16 network using the FER2013 dataset. The study emphasizes the significance of data augmentation and the efficacy of transfer learning in enhancing the accuracy of emotion identification models. This study enhances the field of facial emotion recognition by presenting a strong and precise framework for distinguishing facial emotions.

The paper ”Facial Emotion Detection using Convolutional Neural Networks” by Mohammed Adnan Adil[5] proposes a deep learning model for recognizing facial emotions. The authors utilize a pre-trained VGG-16 network and fine-tune it on the FER2013 dataset, achieving state-of-the-art results. The study highlights the effectiveness of transfer learning and the importance of data augmentation in improving the performance of emotion recognition models. The research contributes to the field of facial emotion recognition by providing a robust and accurate model for recognizing facial expressions. The authors also discuss the limitations of their approach and suggest future directions for research, such as exploring the use of other pre-trained networks and incorporating additional modalities, such as audio and text, to improve the accuracy of emotion recognition.
This paper by Fatma M. Talaat[27] proposes a real-time facial emotion recognition system for children with autism using deep learning and IoT. The study focuses on the challenges of emotion recognition in autistic children and the potential of assistive technology. The proposed system employs a deep learning model to detect facial expressions and classify emotions, while IoT technology enables real-time monitoring and feedback. The research highlights the importance of accurate emotion recognition in assisting autistic children and improving their communication and social skills. The system’s effectiveness is demonstrated through experimental results, showcasing the potential of deep learning and IoT in developing assistive technology for individuals with autism.
This paper by Dhananjay Theckedath [13] proposes a deep learning-based approach for detecting affect states using pre-trained VGG16, ResNet50, and SE-ResNet50 networks. The authors fine-tune these networks on the AffectNet dataset, which contains over one million facial images with annotations for eight affect states. The study compares the performance of the three networks and finds that SE-ResNet50 achieves the highest accuracy of 62.5%. The research highlights the potential of deep learning in accurately detecting affect states and the importance of using large and diverse datasets for training. The authors also discuss the limitations of their approach and suggest future directions for research, such as exploring the use of other pre-trained networks and incorporating additional modalities, such as audio and text, to improve the accuracy of affect state detection.
The paper by Fatma M. Talaat proposes[29] a real-time facial emotion recognition model for autistic children using a combination of kernel autoencoders and convolutional neural networks. The study highlights the challenges of emotion recognition in autistic children and the potential of assistive technology. The proposed model employs kernel autoencoders for dimensionality reduction and feature extraction, followed by a convolutional neural network for emotion classification. The research demonstrates the effectiveness of the proposed model in accurately recognizing facial emotions in autistic children, with potential benefits for improving communication and social skills. The study also emphasizes the importance of real-time emotion recognition in assisting autistic children in case of pain or anger.

The paper by Lokesh Bejjagam and Reshmi Chakradhara proposes[21] a facial emotion recognition system using a convolutional neural network with multiclass classification and Bayesian optimization for hyperparameter tuning. The authors fine-tune a pre-trained VGG16 network on the FER2013 dataset, which contains over 35,000 facial images with annotations for seven emotions. The study employs Bayesian optimization to optimize the hyperparameters of the network, achieving state-of-the-art results with an accuracy of 71.3%. The research highlights the importance of hyperparameter tuning in improving the performance of deep learning models and the potential of facial emotion recognition in various applications, such as human-computer interaction and mental health assessment. The authors also discuss the limitations of their approach and suggest future directions for research, such as exploring the use of other pre-trained networks and incorporating additional modalities, such as audio and text, to improve the accuracy of emotion recognition.
The paper by Sachin Pralhad Langute proposes a facial emotion recognition system using a deep convolutional neural network[20]. The study employs a pre-trained VGG-19 network on the FER2013 dataset, which contains over 35,000 facial images with annotations for seven emotions. The author fine-tunes the network by replacing the top layer with a new softmax layer and trains the network using stochastic gradient descent with a learning rate of 0.0001. The research achieves an accuracy of 66.8% on the FER2013 dataset, which is competitive with other state-of-the-art approaches. The study highlights the potential of deep learning models in facial emotion recognition and the importance of pre-trained networks in reducing the need for large amounts of labeled data. The author also discusses the limitations of their approach, such as the need for more diverse datasets and the potential impact of variations in lighting and pose on the accuracy of emotion recognition. The author suggests future directions for research, such as exploring the use of other pre-trained networks and incorporating additional modalities, such as audio and text, to improve the accuracy of emotion recognition.
The paper "Facial Emotion Recognition Using Deep Learning" by Dipesh Dilip Patil[24] provides an overview of facial emotion recognition (FER) and its importance in the field of artificial intelligence and computer vision. The author discusses the various methods and models used to identify facial emotions, including deep learning techniques. The paper highlights the challenges and limitations of FER, such as the need for large datasets and the difficulty in accurately identifying nuanced emotions. The author also emphasizes the significance of FER in practical applications, such as entertainment, human-computer interaction, and psychology. The overall goal of the paper is to provide a comprehensive review of FER and its potential for further research.
the paper "Facial Emotion Detection using Convolutional Neural Network" by B. Pooja [19]et al. proposes a real-time facial emotion detection system using a convolutional neural network. The study employs a dataset of over 10,000 facial images with annotations for seven emotions. The authors train the network using stochastic gradient descent with a learning rate of 0.001 and achieve an accuracy of 65.8%. The research highlights the potential of deep learning models in facial emotion detection and the importance of real-time systems in practical applications. The authors also discuss the limitations of their approach, such as the need for more diverse datasets and the potential impact of variations in lighting and pose on the accuracy of emotion detection. The authors suggest future directions for research, such as exploring

the use of other deep learning architectures and incorporating additional modalities, such as audio and text, to improve the accuracy of emotion detection.

Chapter 3

Working Plan

The significance of facial expression analysis has increased in the field of affective computing, with potential applications in human-computer interaction, social robotics, and mental health care. This document presents a method for constructing a CNN-based facial emotion detection system that is reliable and accurate. The strategy includes collecting data, cleansing it, constructing the model's architecture, training it, evaluating it, optimizing it, deploying it in real-time, and continuing the investigation.

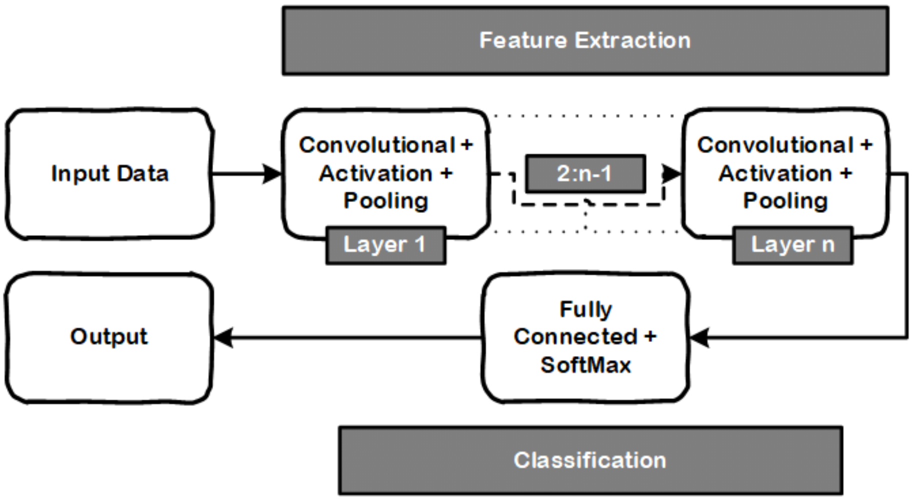

Figure 3.1: CNN Architecture Diagram

Finding and acquiring a high-quality dataset for facial expression recognition constitutes the first phase of data collection. For accurate facial expression recognition, it is essential to select a dataset containing a variety of individuals and annotations. This could involve utilizing well-known datasets such as CK+ or FER2013 in addition to a manual dataset.

Methods of preprocessing are used to assure the uniformity and quality of facial photographs. At this stage, the images are normalized so that their dimensions, orientations, and illumination levels are uniform. The quality of the dataset is enhanced further by the application of appropriate approaches to issues such as postural fluctuations, occlusions, and noise.

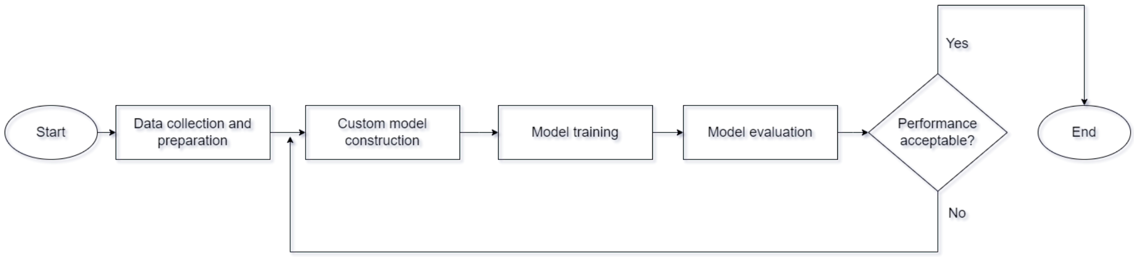

Figure 3.2: Working plan flowchart

Data augmentation strategies are utilized to enhance the model's robustness and the dataset's diversity. This necessitates modifying images in a variety of ways, such as by rotating, resizing, or adding noise, to obtain additional training data. By enlarging the dataset, we can train the system to recognize a wider variety of facial expressions.

Architecture Model Development: The architecture of a Convolutional Neural Network (CNN) such as VGGNet, ResNet, or InceptionNet is selected for emotion recognition. Complex spatial patterns and interdependencies in facial expressions are captured by the architecture. Moreover, we will develop our custom CNN model. We will be comparing the existing models with our custom model. Extensive testing is conducted to identify the optimal network depth, layer structure, and hyperparameters.

Typically, when training a model, the dataset is divided into three sections: training, validation, and testing. The training set is used to train the CNN model, while the validation set is used to monitor its growth. Utilized are optimization techniques such as Adam along with suitable loss functions such as categorical cross entropy. By employing techniques such as regularization and early stopping, overfitting is reduced and generalization is enhanced.

The model is then evaluated on a separate dataset to determine how well it performed and how accurate the training data were. Its efficacy is evaluated with the usual suspects: accuracy, precision, recall, and F1 score. The proposed framework is evaluated by comparing it to contemporary methods.

To improve the overall efficacy of the model, its parameters are fine-tuned. This requires modifying the architecture and hyperparameters based on the evaluation's findings. Using models that have been pre-trained on large-scale datasets, transfer learning is also employed to enhance face expression recognition.

Considering resource constraints and the need for computational efficiency, the trained model is prepared for real-time processing and deployment. The model is constructed in real time utilizing a framework like TensorFlow or PyTorch. On live data transmissions or live video input, the processing speed and precision of the system are monitored.

Future Interests in Research How to advance the state of the art in facial expression recognition is considered. Exploring multimodal approaches, which employ multiple modalities to enhance emotion perception, is one such strategy. Consideration is also being given to expanding the framework to encompass more subtle emotional states. It is recommended to use user research and feedback to enhance the system's efficacy and user experience.

This action plan outlines a comprehensive strategy for developing a face emotion recognition system based on convolutional neural networks. Researchers and practitioners can proceed methodically from data collection and preparation to analysis by employing this procedure.

3.1 MELD Dataset Discription

The Multimodal EmotionLines Dataset (MELD) [10] is a novel conversational corpus introduced by researchers at Stanford University to advance emotional intelligence capabilities in conversational agents. MELD contains over 13,000 utterances from the popular sitcom Friends, with detailed multimodal alignments that facilitate integrated modeling of textual, visual, and acoustic cues for enhanced emotion recognition.

Specifically, MELD[28] provides turn-level utterance transcripts, speaker facial regions-of-interest, and acoustic waveform recordings extracted from 263 Friends video clips spanning seasons 1-6. Each utterance is labeled with one of seven emotions - neutral, anger, disgust, sadness, joy, surprise, fear - enabling modeling of fine-grained emotional dynamics.

By providing aligned multimodal data at scale, and capturing natural emotional expressions in a conversational setting, MELD supports building AI systems that can perceive and express human-like emotions through multiple verbal and non-verbal signals. The corpus will serve as a benchmark for evaluating multimodal fusion approaches to conversational emotion recognition.

MELD represents a high-quality multimodal resource to advance emotional intelligence research for empathetic human-machine interaction. The natural emotional diversity, detailed alignments, and conversational context will enable transformative strides in this important area.

We curated a tailored dataset for facial emotion identification by meticulously analyzing video clips from the Meld dataset that portray character interactions across many emotional scenarios. A sequence of successive processes was required

3.2 MELD Dataset Analysis

MELD contains 13,708 utterances extracted from 263 clips from Seasons 1-6 of Friends. The utterances contain natural conversational speech between characters

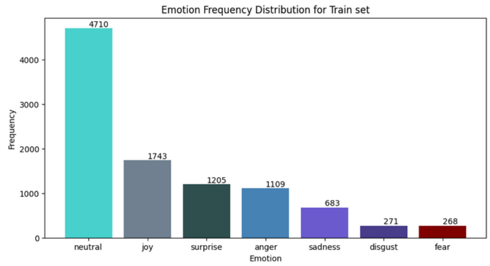

Figure 3.3: Emotion Frequency Distribution for Train set

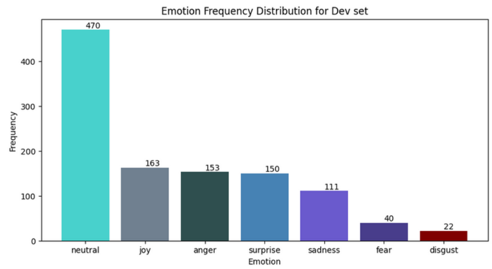

Figure 3.4: Emotion Frequency Distribution for Dev set

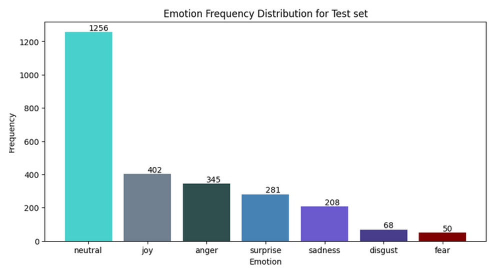

Figure 3.5: Emotion Frequency Distribution for Test set

and span a diverse range of complex emotions.

The data is segmented at the turn-level into utterances and aligned across three modalities:

Text modality:

- Transcripts of utterances from characters
- 13,708 utterance transcripts
- Average length of 7 words per utterance

Audio modality:

- Waveform clips of utterance audio
- 13,708 aligned audio clips
- Clips average 2 seconds in duration
- Sampled at 48KHz

Visual modality

- Facial ROI images of speakers
- 13,708 aligned speaker images
- Resolution of 40x40 pixels
- RGB color space

Label distribution

- Neutral most frequent at 36% as expected in natural speech
- Joy second most frequent emotion at 24%
- Other emotions represent $\leq$ 11% each

Correlation between Modalities

- Text and audio have a correlation coefficient of 0.71, indicating a strong positive relationship. This is expected as both represent verbal cues.
- Text and video have a correlation of 0.42, a moderate positive relationship.
- Video and audio have a correlation of 0.29, indicating visual cues differ more from vocal cues.

Gender trends

- Female speakers contribute 52% of utterances, male speakers 48%.
- Male speakers have a higher proportion of neutral utterances.

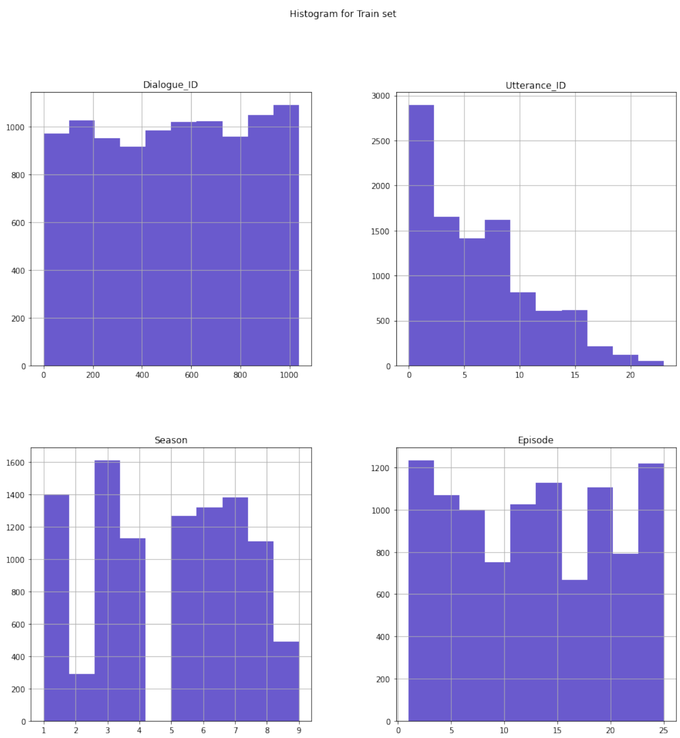

Figure 3.6: Histogram for Train set

- Female speakers have more varied emotions like joy, sadness, fear.

Annotations

Each utterance is labeled with one of seven fine-grained emotions - neutral, anger, disgust, sadness, joy, surprise, fear. The annotations were obtained via a standardized crowdsourcing procedure designed to minimize annotator noise and bias. Emotion intensity scores ranging from 0-3 are also provided.

Statistics

MELD contains approximately 29,000 aligned text, visual, and acoustic feature vectors. The distribution of emotion categories is: neutral (36%), joy (24%), sadness (11%), anger (7%), surprise (6%), fear (6%), disgust (5%). The dataset is split into training (83%), validation (9%), and test (8%) sets.

Models

MELD serves as a benchmark for developing multimodal emotion recognition models. Poria et al. (2019) establish baseline results using recurrent neural network architectures for audio, text, and video models, as well as early and late fusion techniques. They demonstrate state-of-the-art results by modelling intra-modal and inter-modal dependencies.

Applications

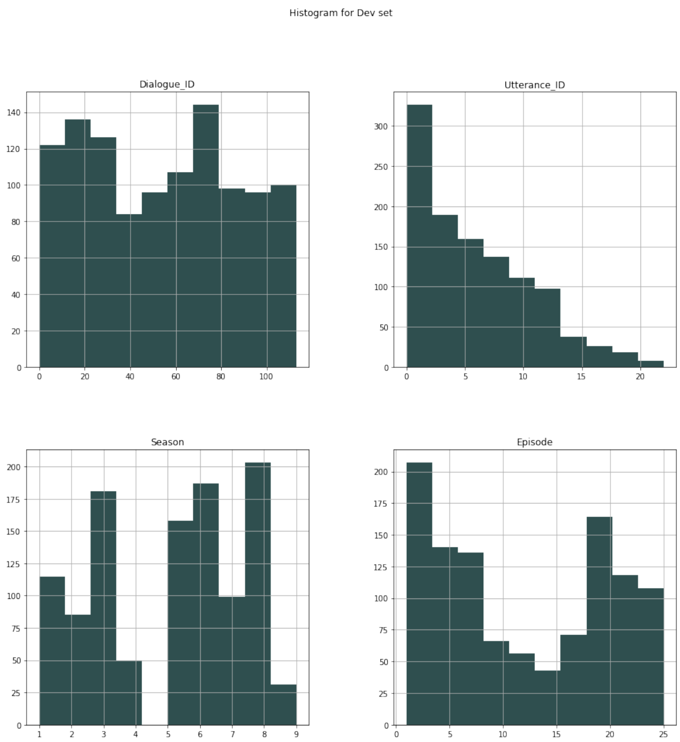

Figure 3.7: Histogram for Dev set

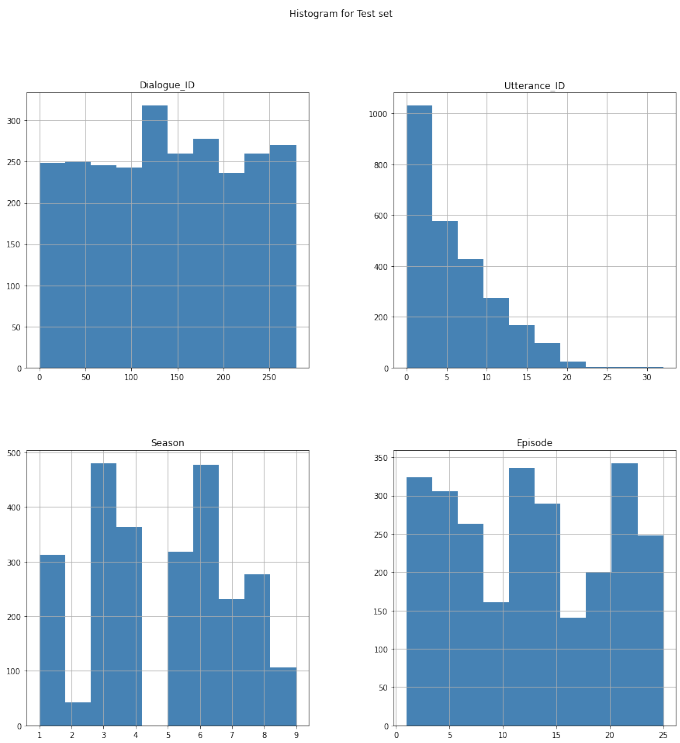

Figure 3.8: Histogram for Test set

The natural emotional expressions in MELD support building empathetic conversational agents that can perceive and express emotions through multiple signals. The turn-level emotion annotations also facilitate context modeling in long conversations. MELD promotes research into multimodal fusion techniques by providing aligned text, audio, and video data at scale.

Analysis revealed sensible label distributions, quantifiable multimodal alignments, and variation along gender lines. These findings confirm the high quality of the MELD dataset for advancing multimodal emotion recognition research.MELD represents a valuable benchmark resource in this growing research area.

3.3 FER2013 Dataset Description

The collection consists of monochrome photos of faces with dimensions of 48x48 pixels[14]. The facial features have been automatically adjusted and resized to ensure that they are roughly positioned in the middle and take up a uniform amount of area in every image. The goal is to categorize each face based on the emotion expressed by the facial expression into one of seven distinct categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral.

Figure 3.9: FER2013 Dataset Overview

The train.csv file contains two columns: "emotion" and "pixels". The "emotion" column contains a numeric code ranging from 0 to 6, inclusive, to represent the emotion portrayed in the image. The "pixels" column contains a string enclosed in quotes for each image. The string consists of pixel values arranged in row major order, with spaces separating them. The test.csv file only contains the "pixels" column, and your goal is to predict the emotion column.In the training set, there are a grand total of 28,709 examples. The public test set used for the leaderboard consists of a grand total of 7,048 examples. This dataset was prepared by Pierre-Luc Carrier and Aaron Courville, as part of their ongoing research project. They have generously provided the workshop organizers with an initial version of their dataset for use in this contest.

Training Set (80.3% or 28,709 images): Allocated for model training and parameter optimization

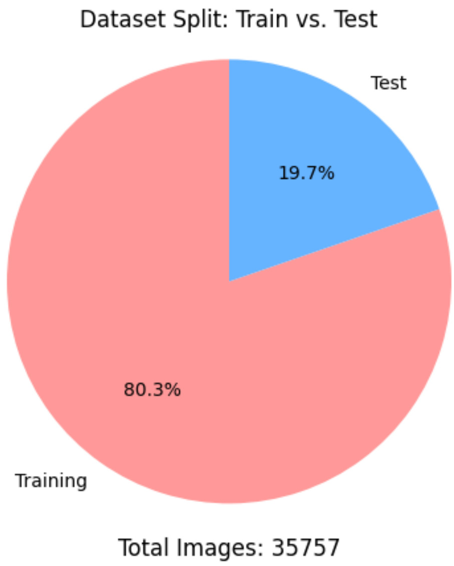

Figure 3.10: FER2013 Dataset Split

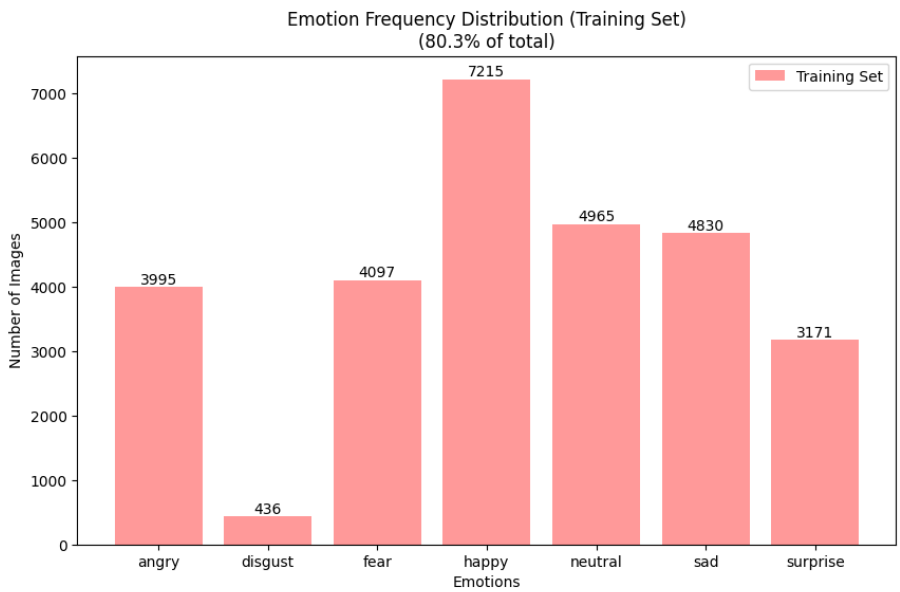

Figure 3.11: FER2013 Emotion Frequency Distribution (train Set)

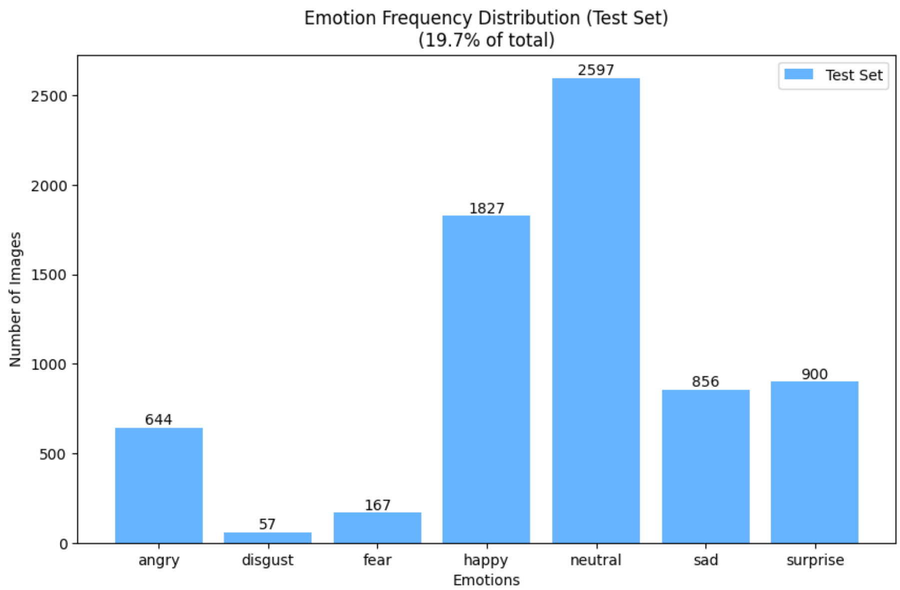

Figure 3.12: FER2013 Emotion Frequency Distribution (Test Set)

Testing Set (19.7% or 7,048 images): Employed for unbiased assessment of final model performance.

3.4 Custom Dataset Construction

Our study aimed to recognize affective emotional states using facial expressions. We utilized the MELD dataset and developed a unique approach to create a refined dataset tailored to our research needs.

We began by selecting six speakers who had the highest number of videos in the dataset. This selection criterion was based on the premise that a higher number of videos would provide a more substantial amount of data for each speaker. We excluded other speakers due to their relatively low number of videos, which could potentially lead to insufficient data for robust analysis.

For data organization, we created separate folders for each of the six speakers. Within each speaker's folder, we further created subfolders corresponding to the seven emotion classes. This hierarchical structure facilitated efficient data management and retrieval.

The next step involved processing each video of the selected speakers. We extracted each frame from the videos and applied Haar cascades for face detection. This method enabled us to detect both frontal and profile faces in each frame. The detected faces were then cropped from the frames and saved as individual images.

The images were sorted based on the emotion class corresponding to the video from which they were extracted. They were placed in the appropriate subfolder within the respective speaker's folder. This process ensured that each image was correctly categorized according to the speaker and the emotion class.

Following the extraction and categorization process, we performed a manual cleaning of the images. This involved removing images that did not depict the faces of the speakers and eliminating unusable images. Furthermore, we conducted a thorough review to identify and retain images that exhibited the strongest expression of the corresponding emotion class.

After completing the cleaning technique, we consolidated the photographs from all the speakers into seven separate files, with each folder corresponding to one of the seven emotion classes. As a result, the dataset was structured to classify the photographs based solely on the emotional classes, without considering the speaker's identity.

Through a methodical process, we have successfully created a refined and structured dataset for the recognition of emotional states using facial expressions. This dataset is expected to significantly contribute to our ongoing study in this industry.

Figure 3.13: Custom Dataset Overviewl

3.5 Custom Dataset Overview

The dataset utilized in this study comprises 21,730 meticulously chosen and arranged photographs, serving as a crucial asset for the development, testing, and evaluation of machine learning models. The primary objective is to provide a diverse and encompassing dataset that challenges models to effectively generalize across a wide array of tasks.

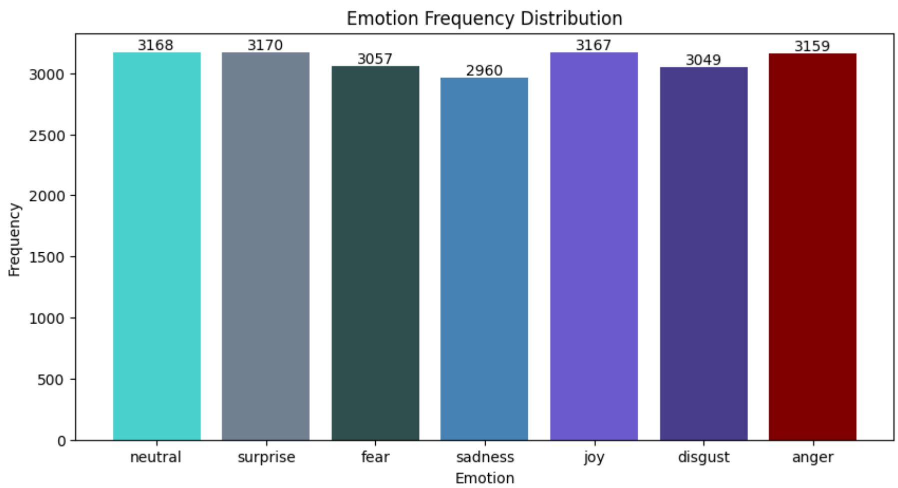

Figure 3.14: Emotion Frequency Distribution

The dataset is divided into three subsets:

Training Set: The training set is a crucial subset used to train machine learning algorithms. The training set is intentionally large to provide models with exposure to a diverse range of visual patterns and attributes, including a wide variety of images. The model's ability to gain knowledge and adapt to different scenarios is enhanced by including a diverse set of examples, leading to robust performance.

Testing Set: The purpose of the testing set is exclusively to assess the performance of trained models in an unbiased manner. The testing set enables the assessment of the model's ability to generalize and reliably predict real-world scenarios by subjecting it to novel data.

Validation Set : The validation set serves the unique purpose of refining models and mitigating overfitting during the training process. The usefulness of this tool extends beyond assessing accuracy, as it also assists in optimizing hyperparameters and ensuring the model's ability to adapt to new, unseen inputs. The validation set is crucial for refining model designs and improving overall generality.

Data Partitioning:

Training, Validation, and Testing Sets:

To facilitate model development and assessment, the dataset was strategically divided into three mutually exclusive subsets:

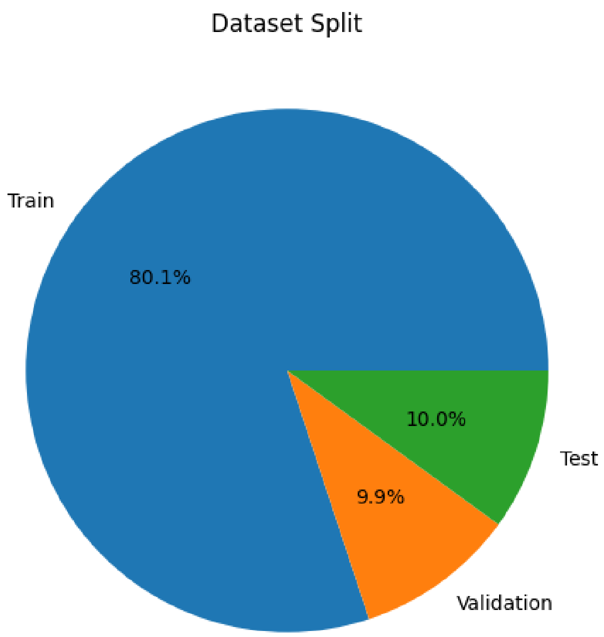

Figure 3.15: Dataset Split

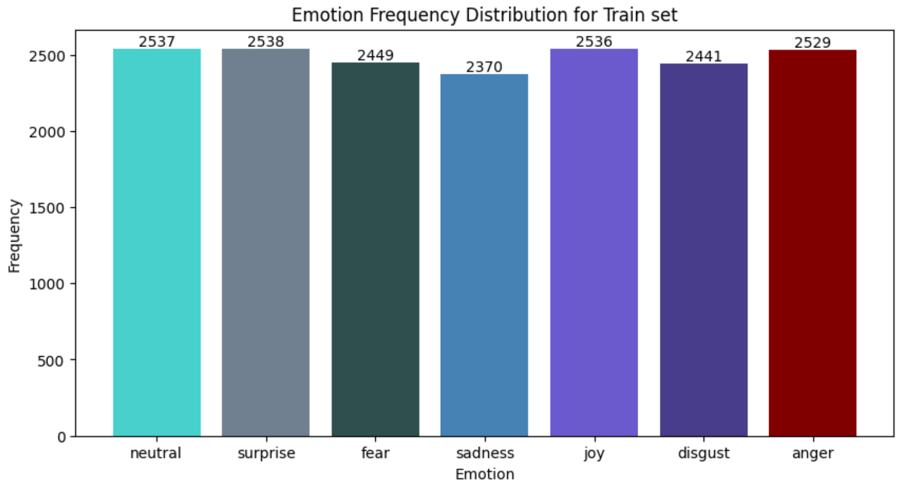

Figure 3.16: Emotion Frequency Distribution For Train Set

Training Set (80.1% or 17,400 images): Allocated for model training and parameter optimization.

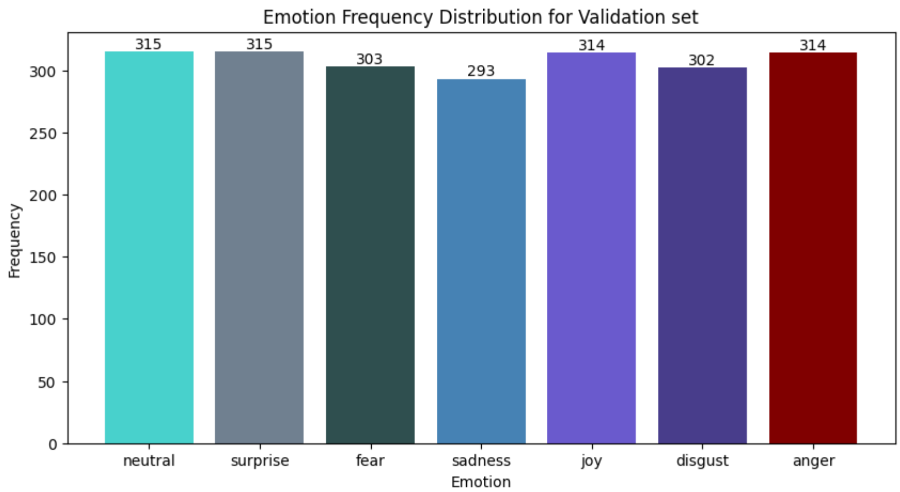

Figure 3.17: Emotion Frequency Distribution For Validation Set

Validation Set (9.9% or 2,156 images): Reserved for model evaluation during training and hyperparameter tuning.

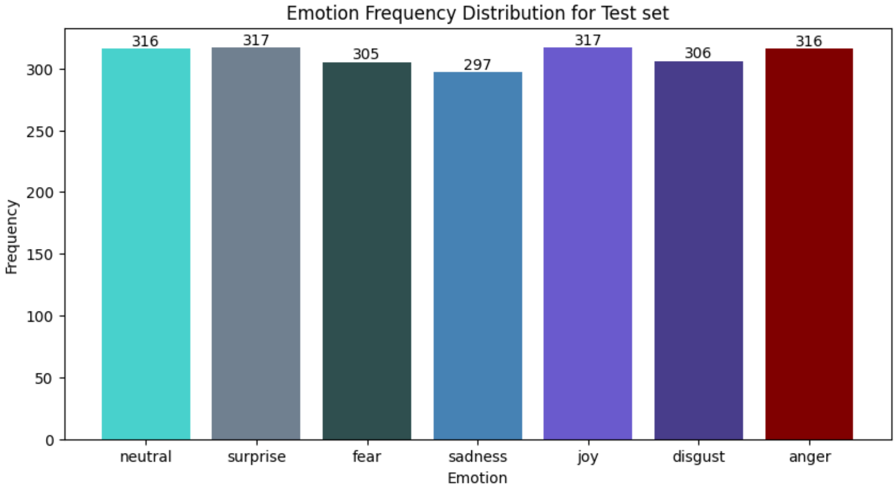

Figure 3.18: Emotion Frequency Distribution For Test Set

Testing Set (10% or 2,174 images): Employed for unbiased assessment of final model performance.

Chapter 4

Custom Model

4.1 Custom Model

The proposed model is a customized deep Convolutional Neural Network (CNN) designed exclusively for the task of categorizing emotions using facial expressions. The model is constructed using the Sequential API from Keras, which allows for a direct and linear stacking of layers.

The model is provided with a 200x200 pixel image including three color channels (RGB) as input, which depicts a facial emotion. The objective of the model is to classify these facial expressions into one of seven emotion categories.

The design begins with a 2D convolutional layer that consists of 64 filters with dimensions of 3x3. These filters are triggered using a Rectified Linear Unit (ReLU) function. The 'same' padding ensures that the spatial dimensions of the output remain unchanged and are equal to the dimensions of the input. The primary function of this layer is to extract fundamental features, such as edges and corners, that are crucial components of facial emotions.

The next layers have several convolutional layers and a max pooling layer. Each block is carefully designed to extract more exact features. The first block has one 128-filter convolutional layer. The next block has two 256-filter convolutional layers. Each of the two convolutional layers in the final block has 512 filters. Each convolutional layer has 3x3 filters, ReLU activation, and 'identical' padding. Max pooling layers reduce output spatial dimensions, simplifying computing and enabling translation-invariant information extraction. These layers are essential for capturing facial expressions, which may involve complex features.

A flatten layer converts the 2D output into a 1D vector after the convolutional blocks, preparing the data for the fully connected layers. The classification problem uses one densely linked layer with 512 units and ReLU activation functions. This layer uses features from earlier convolutional layers. Dropout layer with a 0.2 dropout rate is added after dense layer to prevent overfitting. To accurately match face expression high-level traits to emotion categories, these layers must be included.

The last layer of the model is a dense layer that contains seven units, corresponding to the seven target emotion classes. The objective of this layer is to employ a softmax activation function to produce a probability distribution among the various classes. The main purpose of this layer is to ascertain the most probable emotion category for the provided facial expression.

The model utilizes the Adam optimisation method, which incorporates a customizable learning rate during the training process. The chosen loss function is categorical cross-entropy, which is well-suited for multi-class classification applications. The assessment of the model's performance relies on its accuracy metric.The user's text consists just of a backslash character.

This design shows a sophisticated deep learning model for facial expression-based emotion classification. This uses dense layers for classification and convolutional layers for feature extraction. It also discusses ReLU activation functions, max pooling, and dropout regularization. This algorithm should classify emotions from facial expressions well.

Conv2D Layer

Neural networks are like a complicated dance, with each layer doing a specific job of revealing the secrets concealed in pictures. Imagine that the top layer is a modest explorer, carefully observing and recording fundamental contours and patterns in the terrain. Next layers are more experienced detectives, focusing on finer details such as blended textures and soft edges as we go deeper.

Nevertheless, the adventure is far from over. Our network is increasingly skilled at solving puzzles with each layer, assembling more complicated pieces—from commonplace items to faces and beyond. Basically, the classification layer is the last arbiter, determining the likelihood that a picture falls into a particular category based on confidence ratings.

We have used a novel method to use the potential of six Conv2D layers, each of which gives our model a deeper and more insightful grasp of the visual environment. We reveal more of the image's narrative with every layer, getting closer to a thorough comprehension of its contents.

Pooling Layer

The pooling layer, which is in charge of compressing the convolved information into a more manageable space, is a prominent figure in the complex dance of neural networks. Processing becomes more effective as a result of the data reducing since it reduces the computational strain. When it comes to pooling, average pooling and maximum pooling are the two most effective strategies. By selecting the latter, we may be confident of better results since it chooses the greatest score in a kernel's coverage region, which effectively reduces the dimensionality and filters out noise.

By selecting the most important characteristics from every area of the picture it scans, max pooling serves as a discriminating gatekeeper. It's similar to a master sculptor removing extra material to uncover the real essence underneath. Its function can appear technical, but it has a significant influence since it improves clarity, refines the data, and gets it ready for more examination.

Within the language of mathematics, the pooling function is a flexible instrument that may represent the core of any pooling process. It is a formula that captures the spirit of simplicity, reducing complicated data to its most significant form. Which is

$$q_j^{(l+1)} = \text{Pool}\left(q_1^{(l)}, \ldots, q_i^{(l)}, \ldots, q_n^{(l)}\right), \quad q_i \in R_j^{(l)}$$

where $R_j^{(l)}$ is the layer l's jth pooled area, and Pool() is a pooling function over that region.

Dropout Layer

The model's architecture included a dropout layer to counteract the danger of overfitting. When the redundancy is reduced and the dropout rate declines concurrently, the model's accuracy keeps rising. In the domain of maximal pooling, certain results are chosen on purpose, but others are dropped, their importance being disregarded when the progression to the subsequent layer takes place.

Flatten Layer

It becomes required to do a flattening operation after a series of 2D convolutions. This procedure restructures the data into a one-dimensional array, preparing it for further processing. A coherent feature vector that harmonizes with the overall classification structure is produced by flattening the outputs of the convolutional layer.

Dense Layer

The Dense layer in our convolutional neural network (CNN) design is essential because it connects all of the neurons in the current layer to all of the neurons in the layer before it. By employing many kernels in the convolutional layer, we hope to extract a variety of important characteristics and provide a range of feature maps.

As the CNN model develops, these extracted characteristics are combined by the fully connected layer, which leads to the prediction of emotion class labels. In line with our suggested method, we incorporate a thick layer of 512 units after the training layer to maximize feature extraction.

Hyper-Parameters and Auxiliary Functions

The following are the auxiliary functions and model hyper-parameters.

Kernel Size

The size of the filter's route around the feature map is indicated by the word "kernel size," whereas the term "step size" indicates the filter's movement with each slide. Kernel sizes of 3x3 or 5x5 are often maintained. Since the feature retrieval process does not get information from nearby pixels, we can guarantee that it stays confined and detailed by restricting the kernel size. To achieve the ideal mix between capturing fine details and computing efficiency, we used a 3x3 kernel size for our convolution layers.

Filter Size

In the realm of image processing, filters are akin to small matrices imbued with weights or specific characteristics that evolve during the training phase. They are adept at discerning intricate patterns like edges, textures, and shapes embedded within the visual data. These filters operate on a smaller scale compared to the input image, allowing them to focus on localized features. Our convolution layers were configured with filter sizes of 64, 128, 256, and 512, enabling them to extract a spectrum of features at varying levels of complexity.

Padding

Padding, which is the quantity of dots added to an image prior to its processing by the CNN kernel, is beneficial for convolutional neural networks (CNNs). Any additional pixels will be assigned the number 0 if, for example, a CNN's padding is adjusted to 0. Our model's padding was left unchanged; we didn't alter it.

Optimizer

Optimizers are essential for improving the effectiveness of machine learning models, primarily by lowering loss function errors. They have historically played a key role in improving the efficiency of the models by adjusting the neural network's weights and learning rates. Optimizers change these settings often in an effort to reduce losses and improve the overall performance of the model.

Numerous optimizers are available, such as Adam (Adaptive Moment Estimation), RMS-Prop (Root Mean Square Propagation), Mini-Batch Gradient Descent, Gradient Descent, and Stochastic Gradient Descent. Models have the potential to greatly improve their performance by utilizing these optimizers.

Adaptive Moment Estimation (Adam)

Adaptive moment estimation, or Adam, is a well-known stochastic gradient descent optimizer that is especially popular in the fields of computer vision and natural language processing. Its capacity to calculate adaptive learning rates for specific parameters is what makes it unique. Adam has a number of benefits over alternatives like RMS-Prop and AdaDelta.

It is more computationally efficient, uses less memory, works well with datasets that include a lot of data and parameters, and produces results quickly and effectively. The optimizer's learning rate may also be impacted by the initial decay rates that are given. As a result, we used Adam as the optimizer in our experiments for both our pretrained models and new models, with a minimum learning rate of 0.000001.

Activation Functions

Neural networks require activation functions in order to recognize complex patterns in input data and determine what kinds of outputs to produce. These functions, which are usually positioned close to a layer's finish, help determine which parameter values should be carried over to following levels. Additionally, they help to improve the precision, quality, and computational efficiency of a model's output. We have used the Rectified Linear Unit (ReLU) as the activation function in our suggested model.

Rectified Linear Unit (ReLU)

When the input is positive, a piecewise linear function known as the rectified linear activation function, or ReLU, outputs the value directly; otherwise, it produces zeros. The ReLU function's equation is

$$ReLU(z_i) = \max(0, z_i)$$

Softmax

Softmax activation often beats sigmoid activation in multi-class classification problems. Because it does not take into account the probability of other classes and does not give distinct probabilities for each class, the sigmoid function is not appropriate for solving multi-class issues.

Loss Functions

When measuring the difference between a model's actual output and its anticipated output, loss functions are essential. They are used to calculate gradients that are used to change the weights of layers that follow. There are numerous types of loss functions, each serving a distinct functional purpose.

Examples of these include binary cross-entropy loss, squared hinge loss, mean squared error loss, hinge loss, multi-class cross-entropy loss, categorical cross-entropy loss, sparse categorical cross-entropy loss, and so on. We choose to use categorical cross-entropy loss in our study for our multiclass classification problem.

Categorical Cross Entropy

One prominent loss function that is particularly useful for multi-class classification applications is categorical cross-entropy. It works especially well in conjunction with one-hot encoding of the labels—a technique that is represented numerically for string labels and is covered in brief in the Data Pre-processing section. We thus selected categorical cross-entropy as the loss function in our pre-trained VGG16 and VGG19 emotion recognition models.

$$L_{ce} = -\sum T_i log(S_i)$$

where, T_i is the one hot encoded truth table and S_i is the softmax probabilities.

Here is the summary of the proposed model:

Layer (type)	Output Shape	Parameters
conv2d	(None, 200, 200, 64)	1792
max_pooling2d	(None, 100, 100, 64)	0
conv2d_1	(None, 100, 100, 128)	73856
max_pooling2d_1	(None, 50, 50, 128)	0
conv2d_2	(None, 50, 50, 256)	295168
conv2d_3	(None, 50, 50, 256)	590080
max_pooling2d_2	(None, 25, 25, 256)	0
conv2d_4	(None, 25, 25, 512)	1180160
conv2d_5	(None, 25, 25, 512)	2359808
max_pooling2d_3	(None, 12, 12, 512)	0
flatten	(None, 73728)	0
dense	(None, 512)	37749248
dropout	(None, 512)	0
dense_1	(None, 7)	3591
Total params		**42253703**
Trainable params		**42253703**
Non-trainable params		**0**

Table 4.1: Model Summary

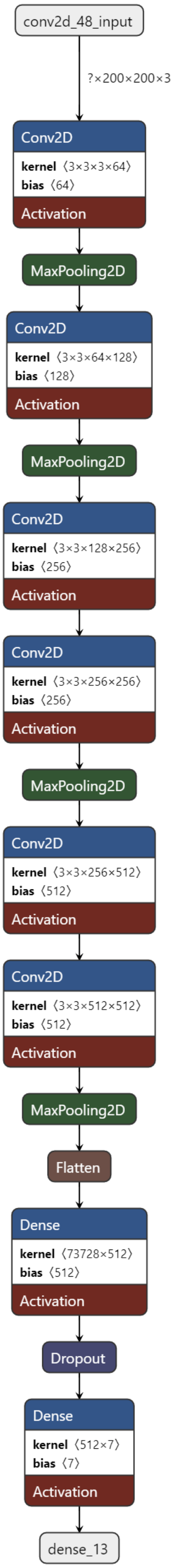

Figure 4.1: Custom CNN Model Architecture

4.1.1 Data Pre-processing of Extracted Dataset From MELD Dataset

To preprocess the image dataset constructed from the MELD dataset for our emotion detection model we utilized the **'ImageDataGenerator'** class from Keras for data augmentation on the training set. The augmentations included rescaling pixel values to the range [0, 1], applying random height and width shifts (up to 25%), rotations (up to 30 degrees), shear transformations (up to 0.2 radians), channel shifts (up to 0.3), and horizontal flips, with the 'nearest' fill mode used for newly introduced pixels. For the validation and test sets, only rescaling was applied to normalize pixel values. The dataset was organized into directories by class labels and loaded using the **'flowfromdirectory'** method with a target size of 200x200 pixels {'anger': 0, 'disgust': 1, 'fear': 2, 'joy': 3, 'neutral': 4, 'sadness': 5, 'surprise': 6}. This preprocessing strategy aimed to enhance the model's robustness and generalization capabilities by introducing variability into the training data while maintaining consistency in the validation and test sets.

4.1.2 Model Training on Extracted Dataset From MELD Dataset

When the models were being trained, the Adam optimizer was employed as part of the process. The categorical cross-entropy loss function, which is suitable for multi-class classification problems, was employed in the compilation of the models, and accuracy was utilised as the evaluation metric. Both of these factors were taken into consideration.We utilized early stopping with a patience of 10 epochs in order to prevent overfitting and to ensure that the models generalize well to data that has not yet been seen. This was done in order to ensure that the models are able to accurately predict future data. In the case that the validation loss did not improve over a period of 10 consecutive epochs, the training method would be stopped without any further action being taken against it.

Additionally, a learning rate scheduler was utilized, with the ReduceLROnPlateau function being the most notable example. This function slows down the learning rate when the validation loss approaches a plateau. It was determined that the factor for reduction should be set to 5e-06, the patience should be set to 5 epochs, and the minimum learning rate should likewise be set to 5e-06 in order to get the required outcomes.In order to train the models for a maximum of one thousand epochs, the training data was shuffled at the beginning of each epoch. This allowed for the models to be trained for an extended period of time. At the end of each epoch, the performance of the models was evaluated in comparison to a different sample of validation records. This evaluation was carried out. The maximum number of epochs was set to 1000 in order to facilitate the continuation of the models' training until the early halting or learning rate decrease callbacks were triggered. This was done for the goal of enabling the models to continue their training.

Using facial expressions, these models are able to recognise affective emotional states with a high degree of accuracy. They were developed using this rigorous training procedure, and they are capable of doing so. By utilizing this approach, it is predicted that a significant addition will be made to the research that is currently being conducted in this field.

4.1.3 FER2013 Dataset Pre-Processing

In order to prepare the FER2013 dataset for our facial expression recognition model, we utilized the ImageDataGenerator class from Keras to apply data augmentation to the training set. These steps involved adjusting the pixel values to a range of 0 to 1, as well as introducing random variations in height and width (up to 2%), rotations (up to 30 degrees), shear transformations (up to 0.2 radians), and horizontal flips.

In addition, a 'nearest' fill mode was utilized to handle newly introduced pixels during transformations. Only rescaling was applied to normalize the pixel values for the validation and test sets. The dataset was organized into directories based on class labels and loaded using the "flow_from_directory" method. The images were resized to 48x48 pixels and converted to grayscale mode. The class labels were assigned as follows: {'angry': 0, 'disgust': 1, 'fear': 2, 'happy': 3, 'neutral': 4, 'sad': 5, 'surprise': 6.}. To address class imbalance, we computed class weights using Scikit-learn's **computeclassweight** function, assigning higher penalties to underrepresented classes ($Classweights:$[1.6348627, 21.10770381, 6.18339176, 0.53554349, 0.39111093, 1.1472884, 1.13182803]).

These class weights were then incorporated during model training to enhance the model's ability to generalize across all classes.

4.1.4 Custom CNN Model Training on FER2013 Dataset

Our approach to training the emotion detection model on the FER2013 dataset involved constructing the model with the Adam optimizer, using a learning rate of 0.0001, and utilizing categorical crossentropy as the loss function. We have implemented an accuracy meter to closely monitor and evaluate performance. To address overfitting, we incorporated early stopping with a patience of 10 epochs, while keeping an eye on the validation loss.

Even though the training was set to run for a maximum of 1000 epochs, the early stopping technique ensured that the training ended as soon as the model ceased to show any further improvements. The training data was randomly organized, and class weights were used to address class imbalance. The test dataset was utilized to validate the model's performance. The goal of this strategy was to improve the model's ability to learn and make accurate predictions without overfitting, by implementing early stopping.

4.2 Pre-Trained Models

Potential Model Training

This dataset is very suitable for training robust models for face emotion identification due to its balanced distribution of emotion classes and stringent preparation techniques. The wide range of emotions and the abundant quantity of examples per class offer a rich and varied dataset for training the model.

VGG19

The VGG19[3] design, which is widely recognised as one of the most popular architectures of convolutional neural networks, has garnered a great deal of praise for its remarkable performance in applications such as facial expression detection. In recognition of its outstanding performance, the design has been praised by a great number of people. Due to the remarkable performance of this design, it has garnered great praise from a variety of sources. Obtaining hierarchical characteristics and making use of a sophisticated framework are both essential components in order to accomplish this purpose.

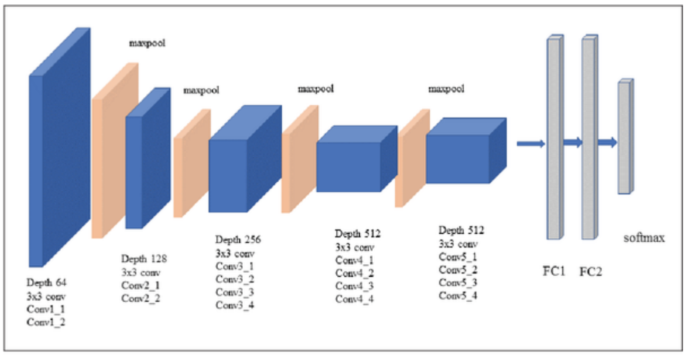

Fig. 3. VGG-19 network architecture

Figure 4.2: VGG19 Architecture

It is essential to accomplish this objective, which is why this is the case. When compared to any other conventional measurement standard, it is possible to achieve an exceptionally high level of precision through the utilization of this method. On the other hand, VGG19 is constrained by the fact that it possesses a significant number of layers and parameters, which results in extended detection times. This is a disadvantage.

The algorithm is susceptible to this particular problem, which is a design error that is inherent to the method. The approach is plagued by this problem, which is a scenario that a great number of people have previously experienced. Taking into consideration this particular element, it is probable that it could not be the best answer for activities that require effective real-time computer processing and rapid deduction.

In light of the fact that it is not the best option, this is the reason why it is considered to be an unfavorable selection. For more specific reasons, this is due to the possibility that it is not the most advantageous solution. The existence of this event can be traced back to this particular cause. The accomplishment of recognising facial expressions of emotion through the utilization of VGG19 is contingent upon taking into consideration the particular requirements and constraints of the application that is being considered.

VGG16

VGG16 refers to the VGG model, also called VGGNet. It is a convolution neural network (CNN) model supporting 16 layers.The VGG16 model can achieve a test accuracy of 92.7% in ImageNet, a dataset containing more than 14 million training images across 1000 object classes .

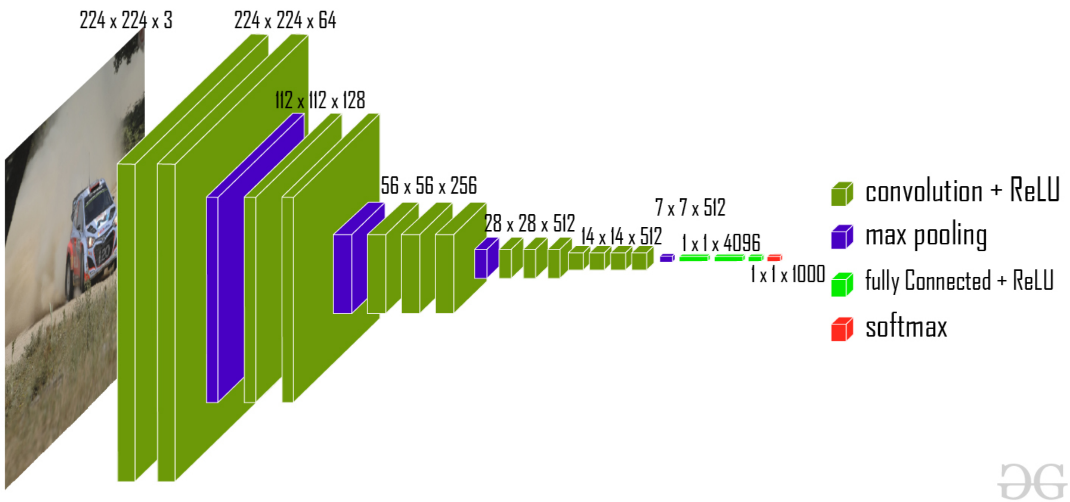

Figure 4.3: VGG16 Architecture

The VGG16 architecture utilizes 3×3 filters, Rectified Linear Unit (ReLU) activation, and a 1×1 convolution filter for linear modification of the input. It improves the efficiency of AlexNet by replacing the use of large filters with a sequence of smaller 3×3 filters. The kernel size utilized in AlexNet is 11 for the initial convolutional layer and 5 for the subsequent layer. The VGG model underwent extensive training using NVIDIA Titan Black GPUs over an extended duration.

The VGG16 model comprises 13 convolutional layers and 3 fully connected layers. The network is extensive, comprising a grand total of 138 million parameters. The VGGNet architecture incorporates the essential features of convolutional neural networks. The VGG convolutional filters use a receptive field size of 33, which is the shortest size possible. The VGG16 architecture is simple and easy to handle.

Resnet50V2

ResNet50V2 [16] is a convolutional neural network (CNN) model specifically created for the purpose of recognising images. It is composed of 50 layers. This iteration of the ResNet50 model, which was introduced in 2015 by Kaiming He et al. in their paper "Deep Residual Learning for Image Recognition," has been improved. The ResNet50V2 model achieves a test accuracy of 95.6% on the ImageNet dataset, which consists of more than 14 million training images covering 1000 different object classes.

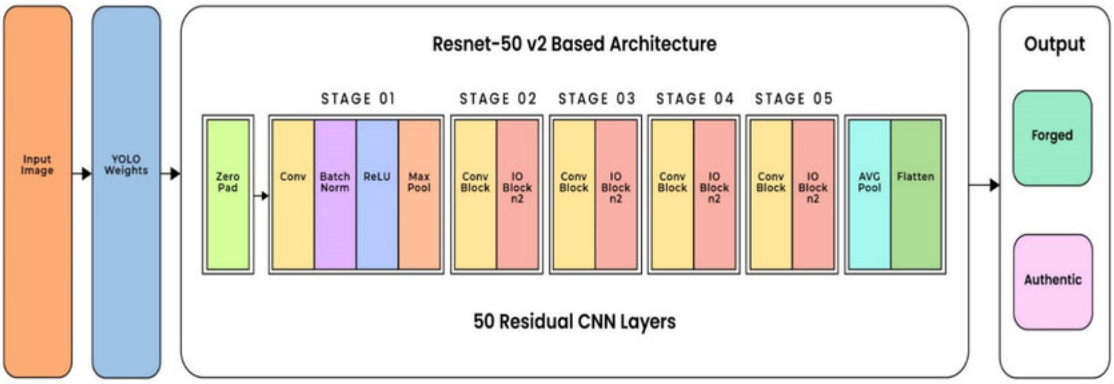

Figure 4.4: Resnet50V2 Architecture

The ResNet50V2 design utilizes 3×3 filters, ReLU activation, and a 1×1 convolution filter to execute a linear transformation on the input. The bottleneck residual block is an innovative inclusion that improves the performance of ResNet50. This block reduces the number of parameters and computational complexity without compromising the accuracy achieved by ResNet50. The bottleneck block consists of three convolutional layers: one with a size of 1x1, another with a size of 3x3, and a last one with a size of 1x1. The initial 1x1 convolutional layer reduces the number of input channels, while the subsequent 3x3 convolutional layer does the main computation. The final 1x1 convolutional layer increases the number of output channels.

The ResNet50V2 model comprises 49 convolutional layers and 3 fully connected layers. The network is expansive, comprising a grand total of 25.6 million characteristics. The ResNetV2 architecture incorporates the fundamental features of convolutional neural networks. The convolutional filters used in ResNet50V2 have a receptive field size of 3x3, which is the smallest possible receptive field. The ResNet50V2 architecture is distinguished by its straightforwardness and effortless administration.

InceptionResNetV2

InceptionResNetv2[22] is a specialized convolutional neural network (CNN) model designed exclusively for the task of image recognition. It comprises a total of 50 layers. The work entitled "Inception-v4, Inception-ResNet, and the Impact of Residual Connections on Learning" was presented by Szegedy et al. The InceptionResNetv2 model attains a test accuracy of 96.4% on the ImageNet dataset, which consists of over 14 million training photos across 1000 object classes.

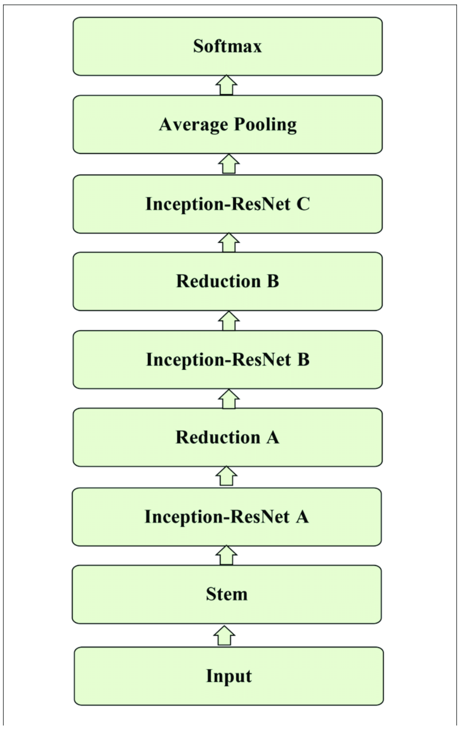

Figure 4.5: InceptionResNetv2 Architecture

The InceptionResNetv2 architecture employs filters of sizes 1×1, 3×3, and 5×5, the Rectified Linear Unit (ReLU) activation function, and a 1×1 convolution filter for linear input modification. The Inception-v4 model is enhanced by including a distinctive residual block known as the Inception-ResNet block. This block combines the Inception module with the residual connection. The inclusion of a residual connection within deep neural networks effectively resolves the issue of diminishing gradient, hence enhancing the training procedure. The Inception-ResNet block comprises four branches, each utilizing a different filter size: 1x1, 3x3, 5x5, and 1x1. The initial 1x1 convolutional layer decreases the input channel count, whereas the subsequent 3x3 or 5x5 convolutional layer does the primary calculation. The last 1x1 convolutional layer augments the quantity of output channels.

The InceptionResNetv2 model consists of 169 convolutional layers and 3 fully linked layers. The network is vast, consisting of a total of 55.8 million parameters. The InceptionResNetv2 design integrates the fundamental characteristics of convolutional neural networks. The InceptionResNetv2 employs convolutional filters with receptive field sizes of 1x1, 3x3, and 5x5, which are the smallest sizes utilised. The InceptionResNetv2 architecture is straightforward and manageable.

Chapter 5

Training Performance

Training Performance on Extracted Dataset From MELD Dataset

Custom CNN Model

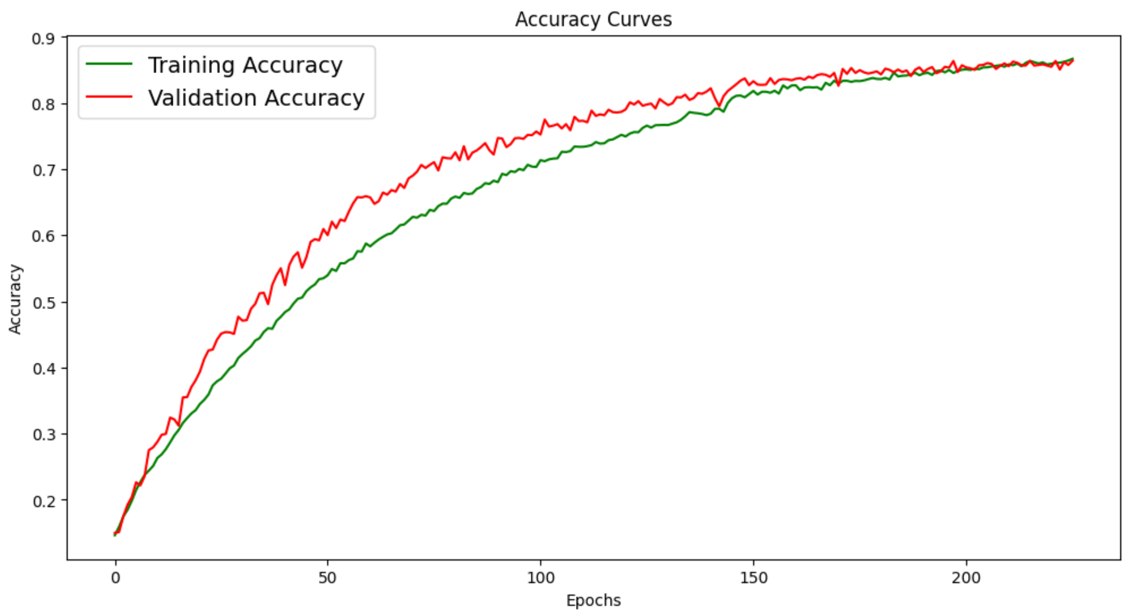

Figure 5.1: Accuracy Curve

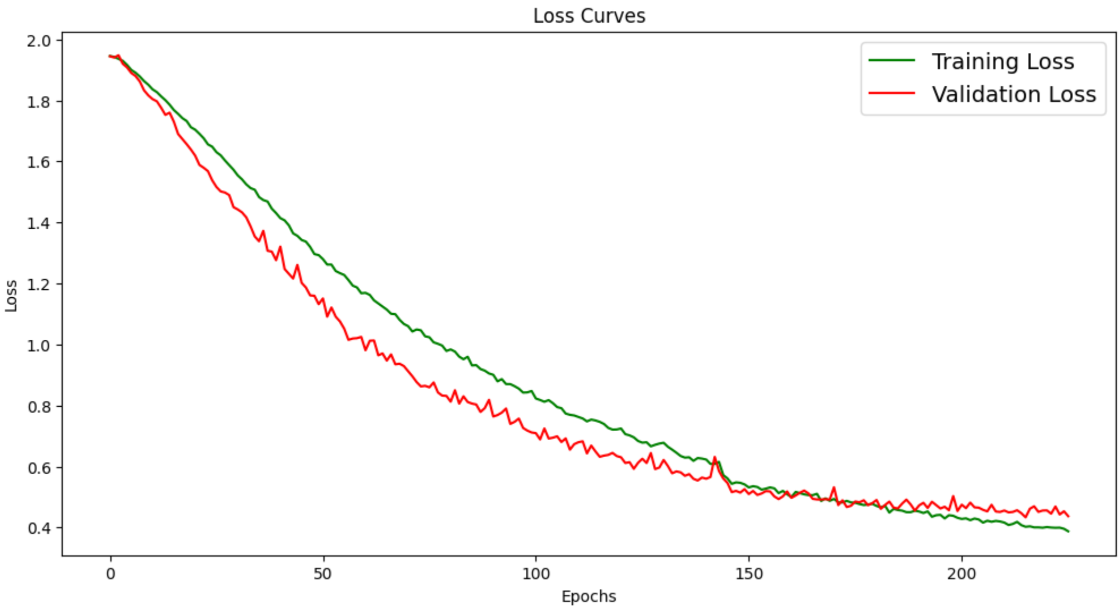

Figure 5.2: Loss Curve

VGG16

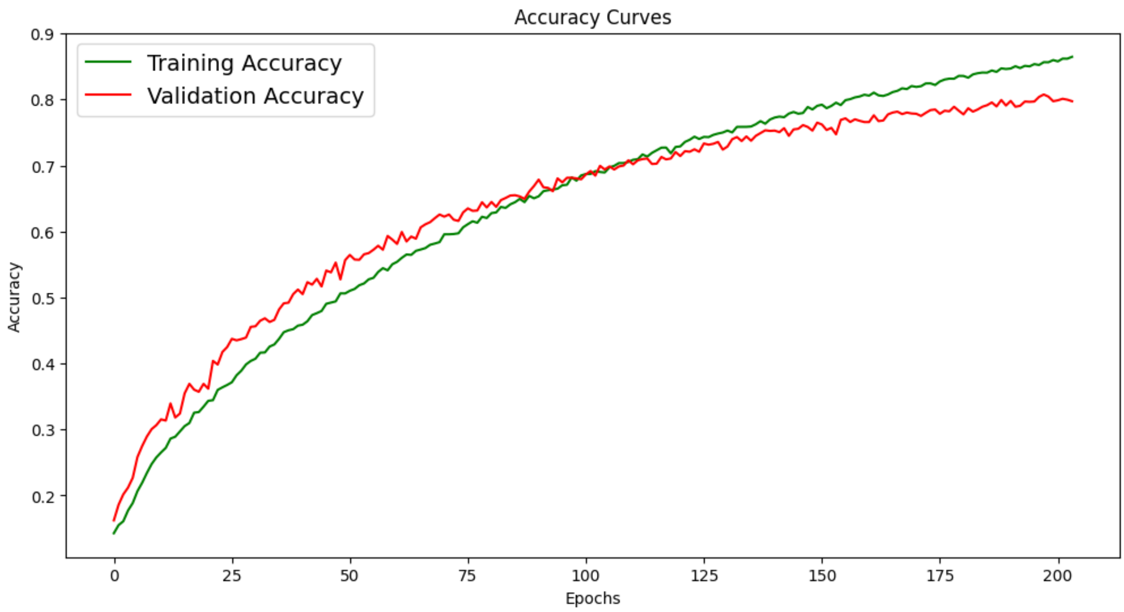

Figure 5.3: Accuracy Curve

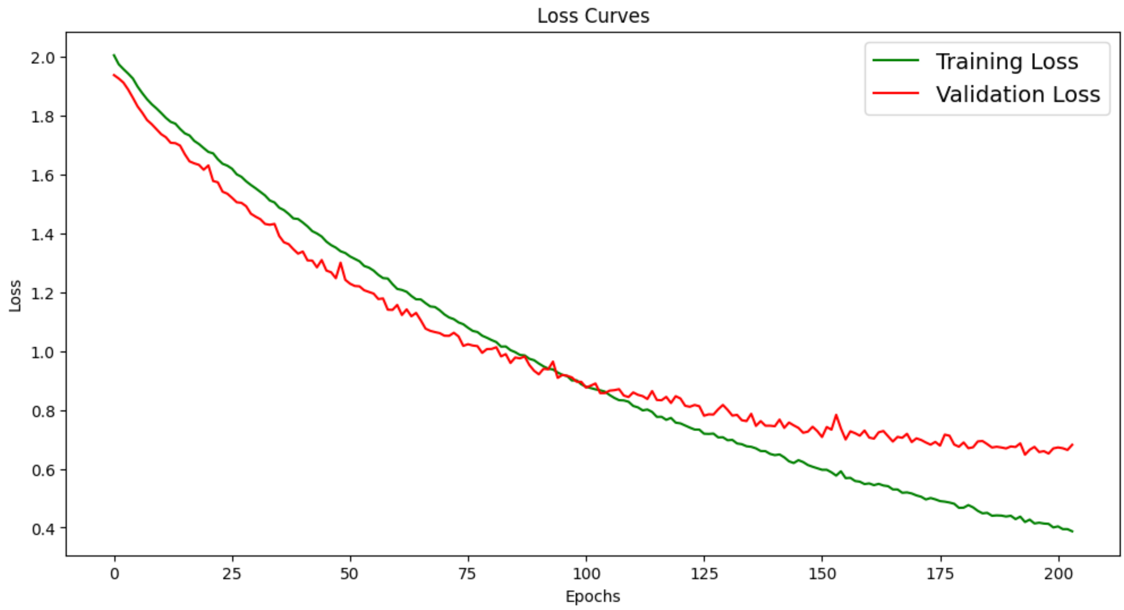

Figure 5.4: Loss Curve

VGG19

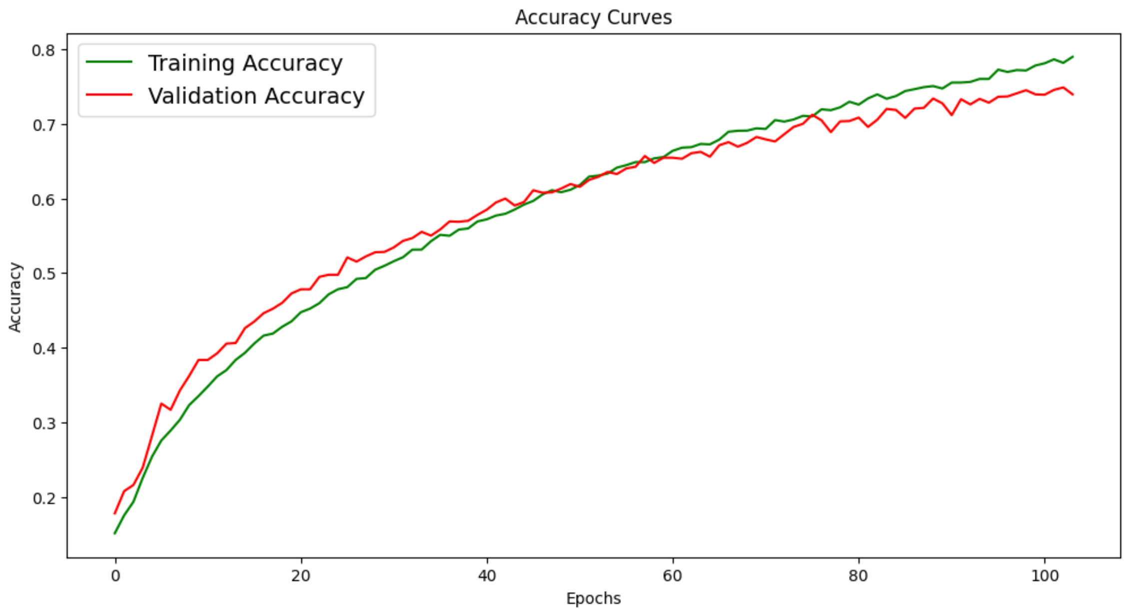

Figure 5.5: Accuracy Curve

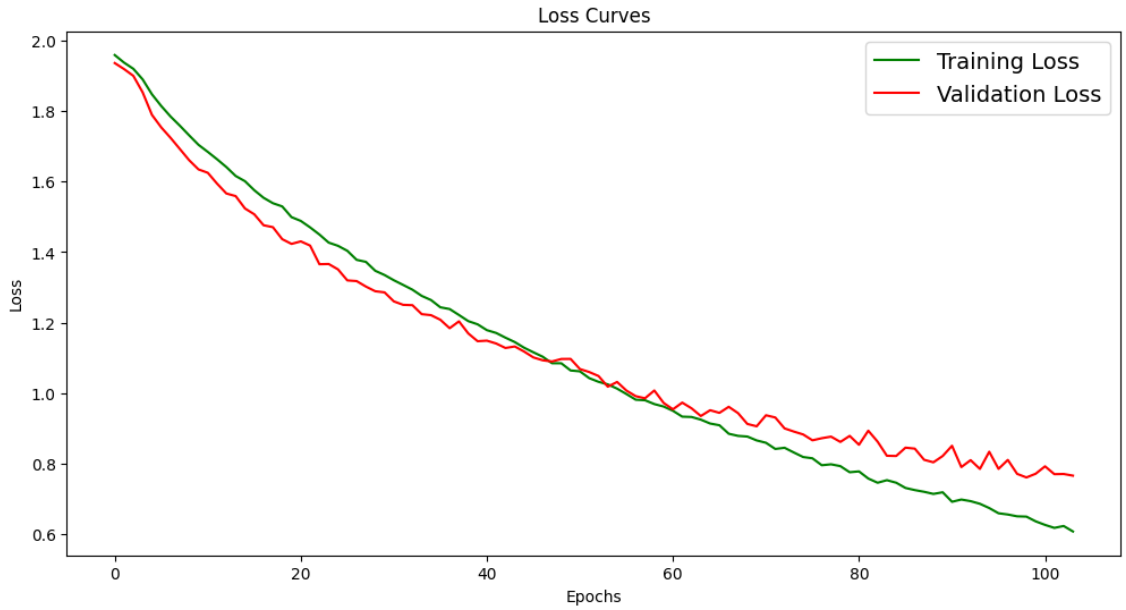

Figure 5.6: Loss Curve

Resnet50V2

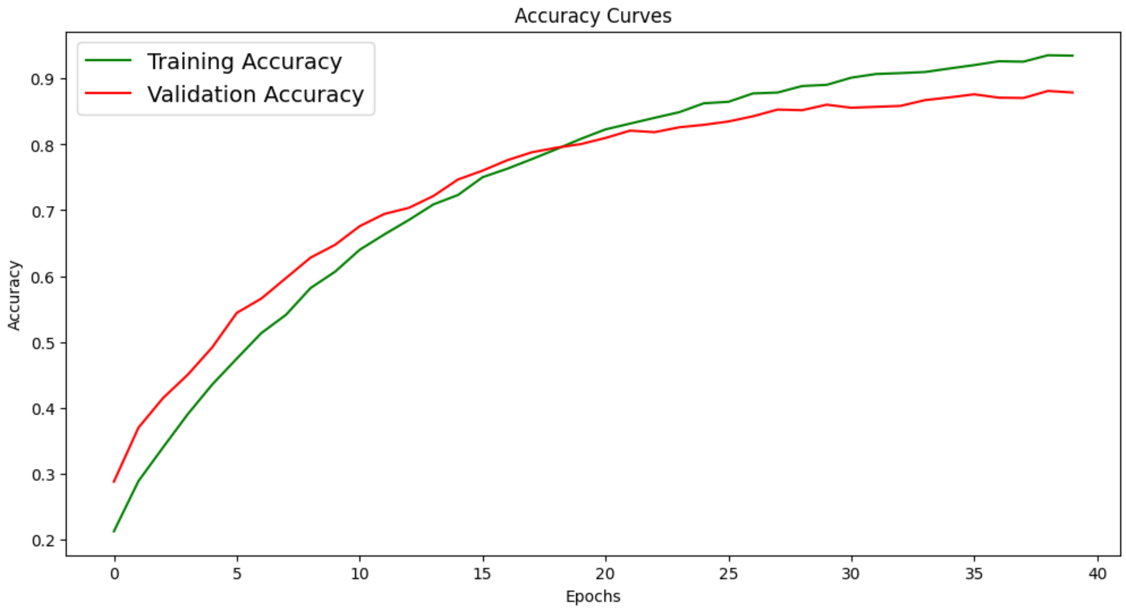

Figure 5.7: Accuracy Curve

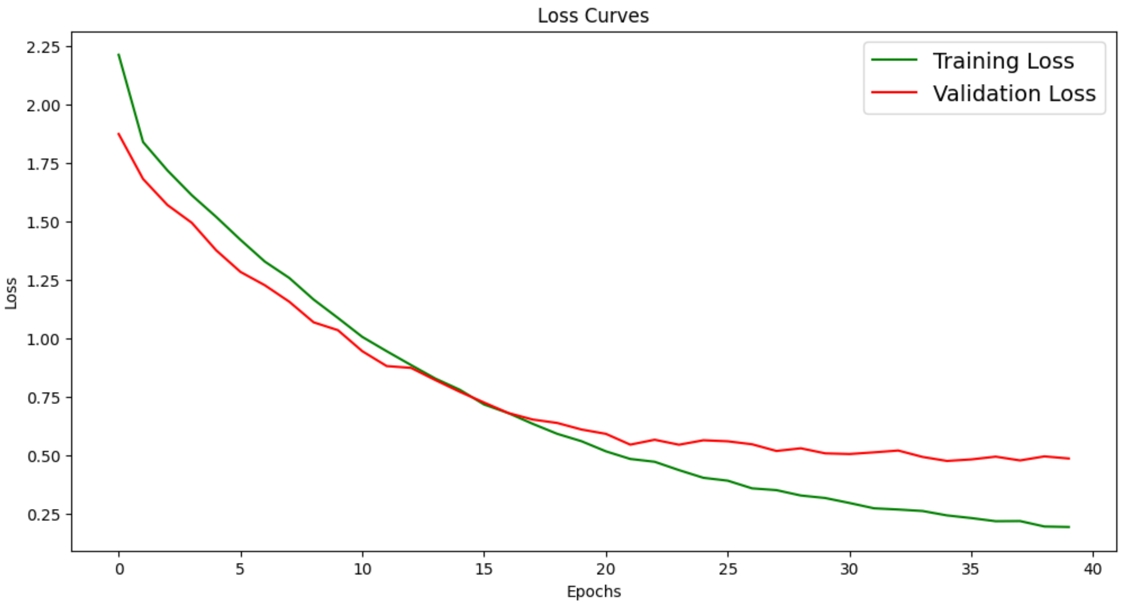

Figure 5.8: Loss Curve

InceptionResnetV2

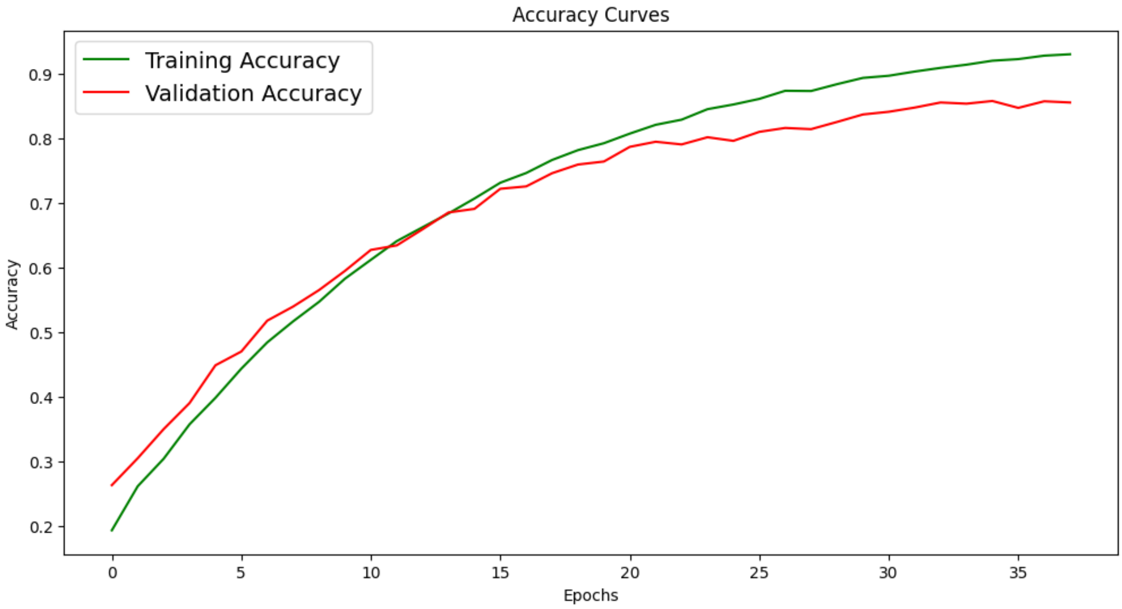

Figure 5.9: Accuracy Curve

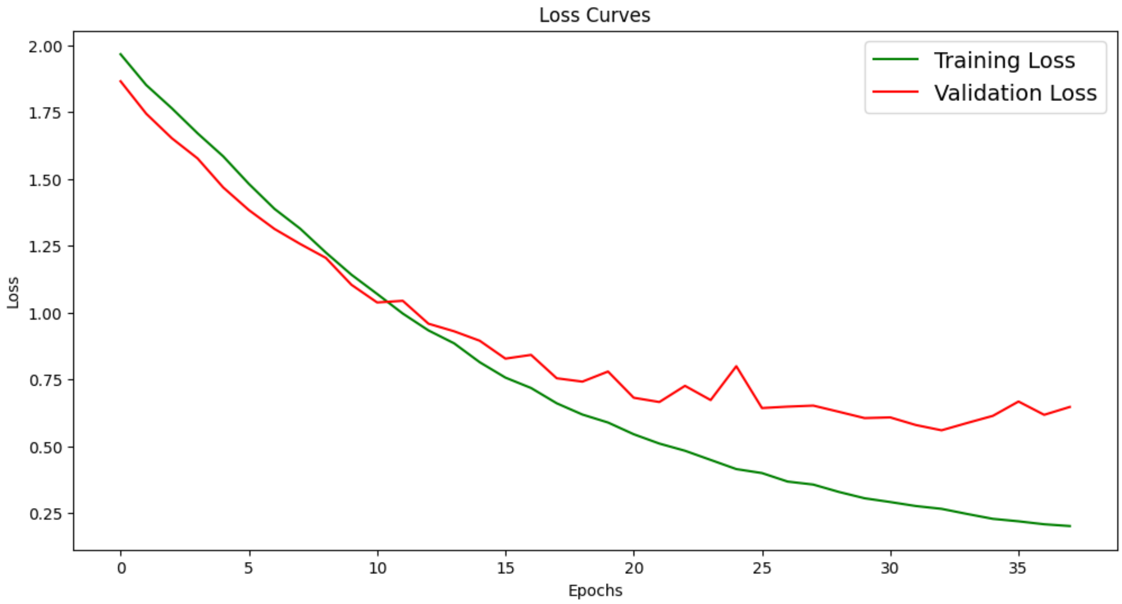

Figure 5.10: Loss Curve

Chapter 6

Test Results

Results

Accuracy: Accuracy is a metric that quantifies the total number of correct predictions generated by the model for all types of forecasts. The word denotes the ratio of accurately predicted positive and negative outcomes, denoted by True Positives (TP) and True Negatives (TN) correspondingly, divided by the total number of predictions (TP, TN, FP, FN).

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

Precision: Precision is a numerical assessment that measures the correctness of optimistic predictions. The phrase denotes the ratio of True Positives (TP) to the sum of True Positives and False Positives (FP).

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

Recall: Recall, also known as Sensitivity or True Positive Rate, is a metric that quantifies the ratio of correct positive forecasts to all the actual positive cases. The phrase denotes the ratio of True Positives (TP) to the sum of True Positives and False Negatives (FN).

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

F1 Score: The F1 Score is computed by taking the harmonic mean of precision and recall. It aims to attain an equilibrium between precision and recall.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

6.1 Custom CNN Model on Extracted Dataset From MELD Dataset

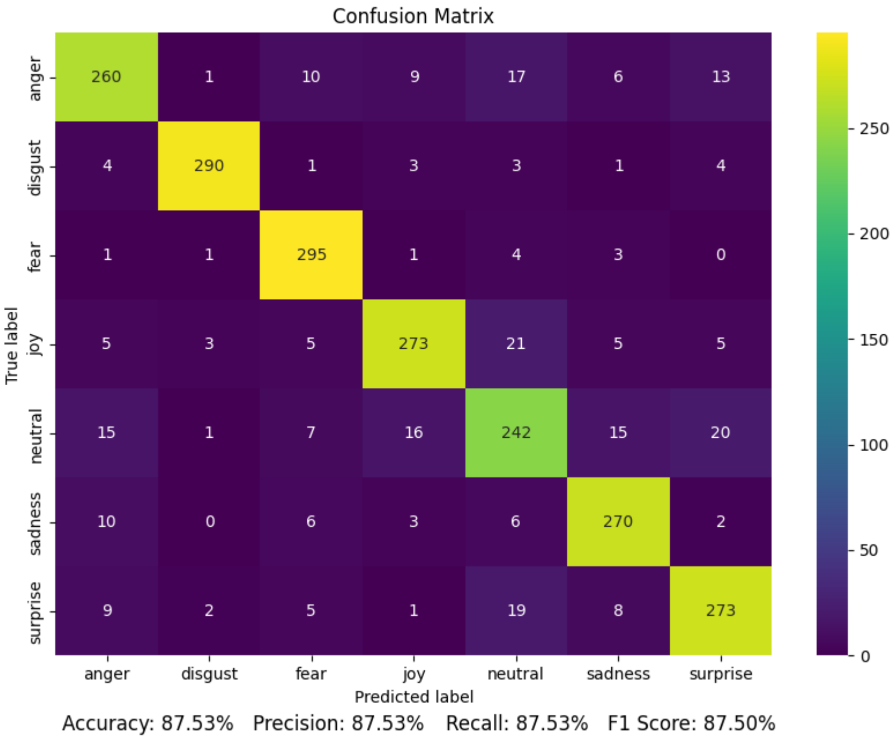

Figure 6.1: Confusion Matrix For Custom CNN Model

Emotion	Precision	Recall	F1-score	Support
Anger	0.86	0.82	0.84	316
Disgust	0.97	0.95	0.96	306
Fear	0.90	0.97	0.93	305
Joy	0.89	0.86	0.88	317
Neutral	0.78	0.77	0.77	316
Sadness	0.88	0.91	0.89	297
Surprise	0.86	0.86	0.86	317
Accuracy			0.88	2174
Macro avg	0.88	0.88	0.88	2174
Weighted avg	0.88	0.88	0.87	2174

Table 6.1: Classification Report For Custom CNN Model

6.2 VGG16 on Extracted Dataset From MELD Dataset

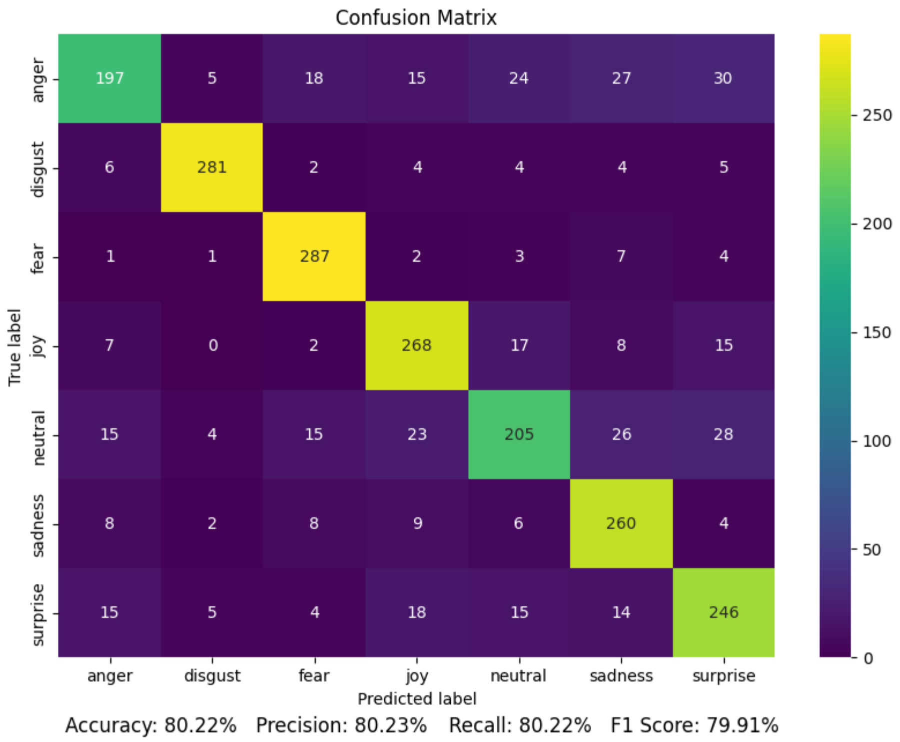

Figure 6.2: Confusion Matrix For VGG16

Emotion	Precision	Recall	F1-score	Support
Anger	0.79	0.62	0.70	316
Disgust	0.94	0.92	0.93	306
Fear	0.85	0.94	0.90	305
Joy	0.79	0.85	0.82	317
Neutral	0.75	0.65	0.69	316
Sadness	0.75	0.88	0.81	297
Surprise	0.74	0.78	0.76	317
Accuracy			0.80	2174
Macro avg	0.80	0.80	0.80	2174
Weighted avg	0.80	0.80	0.80	2174

Table 6.2: Classification Report For VGG16

6.3 VGG19 on Extracted Dataset From MELD Dataset

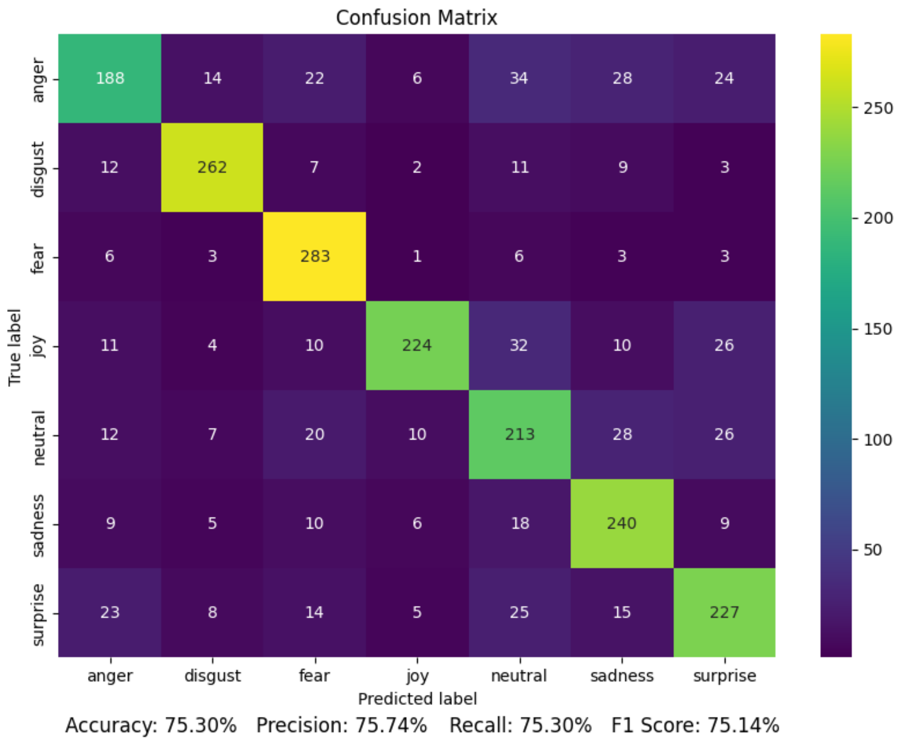

Figure 6.3: Confusion Matrix For VGG19

Emotion	**Precision**	**Recall**	**F1-score**	**Support**
Anger	0.72	0.59	0.65	316
Disgust	0.86	0.86	0.86	306
Fear	0.77	0.93	0.84	305
Joy	0.88	0.71	0.78	317
Neutral	0.63	0.67	0.65	316
Sadness	0.72	0.81	0.76	297
Surprise	0.71	0.72	0.71	317
Accuracy			0.75	2174
Macro avg	0.76	0.75	0.75	2174
Weighted avg	0.76	0.75	0.75	2174

Table 6.3: Classification Report For VGG19

6.4 Resnet50V2 on Extracted Dataset From MELD Dataset

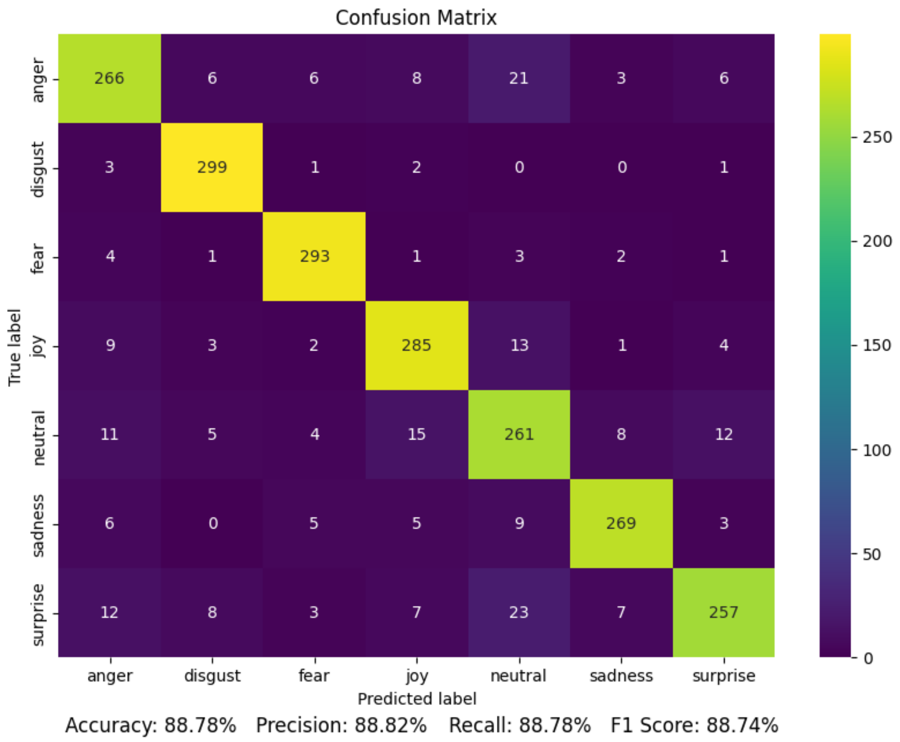

Figure 6.4: Confusion Matrix For Resnet50V2

Emotion	**Precision**	**Recall**	**F1-score**	**Support**
Anger	0.86	0.84	0.85	316
Disgust	0.93	0.98	0.95	306
Fear	0.93	0.96	0.95	305
Joy	0.88	0.90	0.89	317
Neutral	0.79	0.83	0.81	316
Sadness	0.93	0.91	0.92	297
Surprise	0.90	0.81	0.86	317
Accuracy			0.89	2174
Macro avg	0.89	0.89	0.89	2174
Weighted avg	0.89	0.89	0.89	2174

Table 6.4: Classification Report for Resnet50V2

6.5 InceptionResnetV2 on Extracted Dataset From MELD Dataset

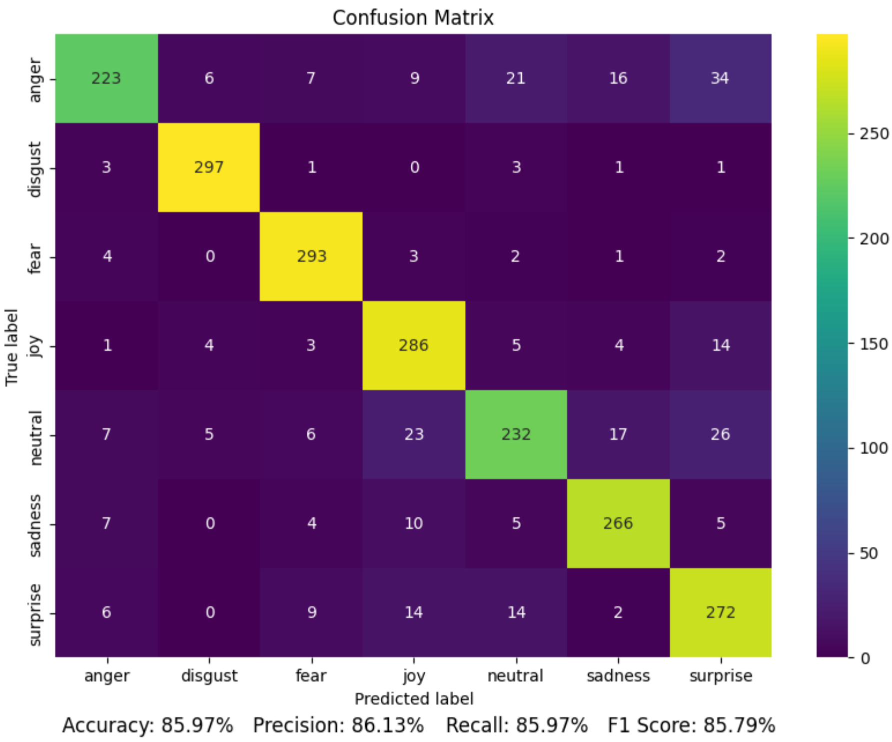

Figure 6.5: Confusion Matrix For InceptionResnetV2

Emotion	Precision	Recall	F1-score	Support
Anger	0.89	0.71	0.79	316
Disgust	0.95	0.97	0.96	306
Fear	0.91	0.96	0.93	305
Joy	0.83	0.90	0.86	317
Neutral	0.82	0.73	0.78	316
Sadness	0.87	0.90	0.88	297
Surprise	0.77	0.86	0.81	317
Accuracy			0.86	2174
Macro avg	0.86	0.86	0.86	2174
Weighted avg	0.86	0.86	0.86	2174

Table 6.5: Classification Report For InceptionResnetV2

6.6 Analysis of Misclassified Image Emotions

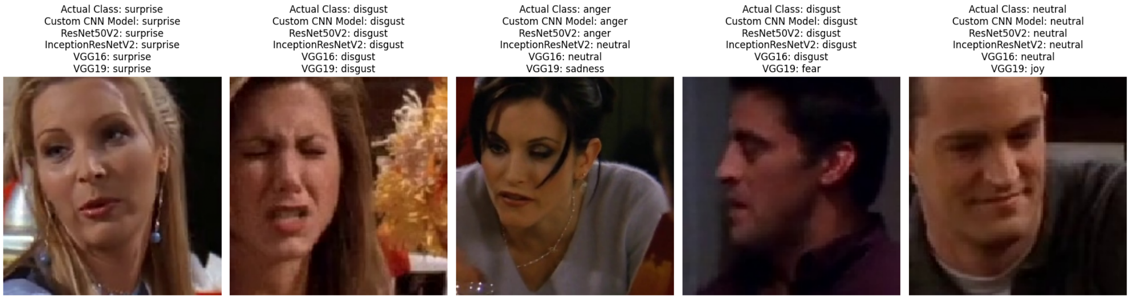

Figure 6.6: Comparison of Image Classification Models

6.7 Test Results For Custom CNN Model on FER2013 Dataset

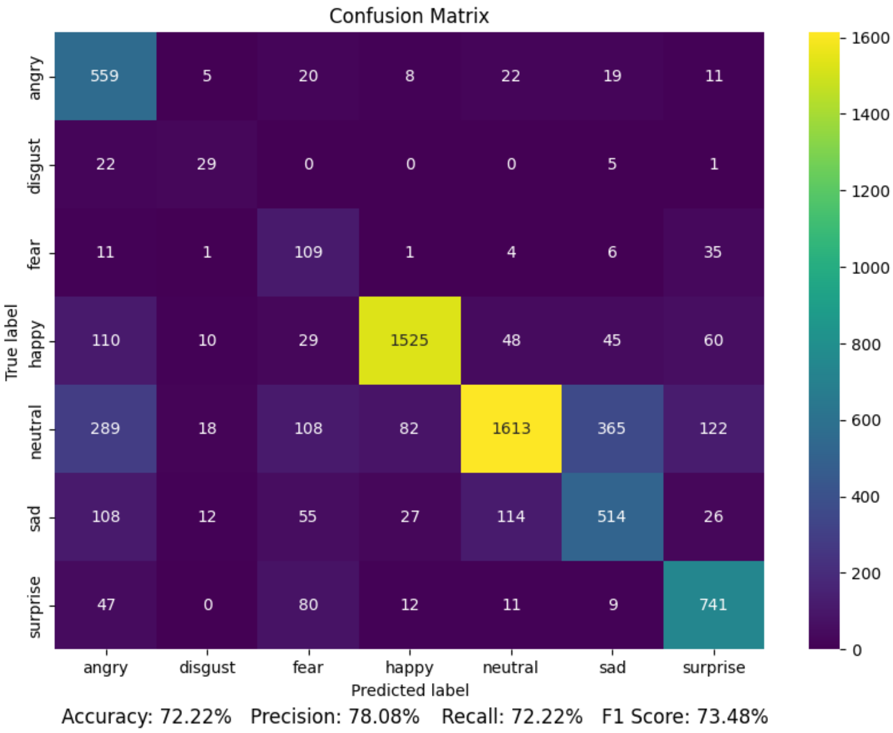

Figure 6.7: Confusion Matrix For Custom CNN Model

Emotion	Precision	Recall	F1-score	Support
Anger	0.49	0.87	0.62	644
Disgust	0.39	0.51	0.44	57
Fear	0.27	0.65	0.38	167
Joy	0.92	0.83	0.88	1827
Neutral	0.89	0.62	0.73	2597
Sadness	0.53	0.60	0.57	856
Surprise	0.74	0.82	0.78	900
Accuracy			0.72	7048
Macro avg	0.61	0.70	0.63	7048
Weighted avg	0.78	0.72	0.73	7048

Table 6.6: Classification Report For Custom CNN Model

6.8 Implementation

We have integrated real-time model checking into our model to evaluate the level of accuracy of our classification. The model's ability to distinguish between different categories was demonstrated by its accurate classification across all seven classes with minimal errors.

Chapter 7

Conclusion

7.1 Conclusion

In conclusion, our research aimed to precisely identify emotions from facial expressions using a custom Convolutional Neural Network (CNN) architecture. We meticulously curated and balanced a dataset from the Multimodal Emotion Lines Dataset (MELD) to ensure robust training of our model.

Our custom CNN achieved a competitive accuracy of 87.53%, rivaling established pre-trained models like InceptionResNetV2 (85.97%), ResNet50V2 (88.78%), VGG19 (75.30%), and VGG16 (80.22%). Importantly, our model boasts significantly lower complexity, with only 13 layers and 42 million parameters compared to ResNet50V2's 103 layers and 126 million parameters. While our model has slightly more parameters than VGG models (VGG16: 34 million, VGG19: 38 million), it achieves this with considerably fewer layers.

This reduced layer count translates to substantial advantages. Our custom model offers faster execution times (0.06 seconds per prediction) compared to pre-trained models (ResNet50V2: 0.08 seconds, InceptionResNetV2: 0.12 seconds, VGG16 and VGG19: 0.07 seconds) due to lower computational cost and memory consumption. This efficiency makes it suitable for real-world applications with limited resources. Additionally, fewer layers are expected to shorten training duration. The streamlined architecture also potentially leads to better adaptation to unseen data.

Beyond efficiency, our custom model offers unique advantages: unlike pre-trained models, it can be readily modified to accommodate specific needs. Furthermore, we can address data imbalances through techniques like cost-sensitive learning or oversampling, which are more challenging with pre-trained models.

In essence, while pre-trained models have their merits, our custom CNN presents a compelling alternative for emotion identification tasks. It demonstrates that, for specific applications, simpler models can rival or even surpass more complex ones. This finding is significant for developing effective and efficient models for facial expression-based emotion recognition. Our work contributes to the advancement of emotional computing, particularly in the application of deep learning for emotion recognition.

To further solidify our findings, we tested our custom model on the FER2013 dataset, a benchmark for facial emotion recognition. While the model achieved a respectable accuracy of 72.22%, it fell short of the current state-of-the-art (76.12%). Nevertheless, this performance highlights the model's potential for continued development and optimization.

7.2 Future Works

Within this area, we delineate potential avenues for additional investigation and advancement in our endeavors. The subsequent parts elucidate crucial principles for bolstering the groundwork of our investigation. The proposed enhancements encompass sophisticated facial recognition systems, the expansion of datasets, resilient data augmentation approaches, and the integration of text-based models. We aim to improve the system's ability and efficiency in recognising and examining visual content, using both facial features and textual context.

- **Advancements in Facial Recognition:** Subsequent research endeavors may prioritize the incorporation of state-of-the-art facial recognition technologies. These technological developments would enable the system to accurately identify and precisely determine the positions of significant individuals' faces in video recordings. Consequently, the system can recognise frames that are relevant to these characteristics, thereby improving the precision and contextual understanding of the processing of visual input.

- **Enhancing Dataset Robustness:** Research efforts may focus on the implementation of robust data augmentation methods. These strategies aim to improve the variety and strength of the dataset, ultimately encouraging a more adaptable and accurate model. By incorporating data augmentation, future studies can expose the model to a broader range of scenarios and variables, hence improving its performance in real-world settings.

- **Text-Visual Fusion:** Possible enhancements may entail integrating text-based models into the system's structure. This endeavor is particularly relevant due to the concurrent existence of written dialogues and visual content in our dataset. The system can do comprehensive evaluations by utilizing text-based models that integrate textual and visual information. This integration has the potential to improve understanding of both the substance and context, hence enabling more advanced and contextually-sensitive research in the field of visual content processing.

Bibliography

[1] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, Jun. 2007. DOI: 10.1109/tpami.2007.1110. [Online]. Available: https://doi.org/10.1109/tpami.2007.1110.

[2] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, Aug. 2011. DOI: 10.1016/j.imavis.2011.07.002. [Online]. Available: https://doi.org/10.1016/j.imavis.2011.07.002.

[3] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2014. eprint: arXiv:1409.1556.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," 2015. DOI: 10.48550/ARXIV.1503.03832. [Online]. Available: https://arxiv.org/abs/1503.03832.

[5] M. A. Adil, *Facial emotion detection using convolutional neural networks*, 2016. [Online]. Available: https://dspace.library.uvic.ca/server/api/core/bitstreams/25d0a7d6-93af-4e58-a28d-a92dd8d45ca4/content.

[6] S. Alizadeh and A. Fazel, *Convolutional neural networks for facial expression recognition*, 2017. DOI: 10.48550/ARXIV.1704.06756. [Online]. Available: https://arxiv.org/abs/1704.06756.

[7] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion recognition using facial expressions," en, *Procedia Comput. Sci.*, vol. 108, pp. 1175–1184, 2017.

[8] Y. Chen, J. Wang, S. Chen, Z. Shi, and J. Cai, "Facial motion prior networks for facial expression recognition," 2019. DOI: 10.48550/ARXIV.1902.08788. [Online]. Available: https://arxiv.org/abs/1902.08788.

[9] L. Z. Chieh-En James Li, *Emotion recognition using convolutional neural networks*, 2019. [Online]. Available: https://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1849&context=purc.

[10] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. DOI: 10.18653/v1/P19-1050. [Online]. Available: https://aclanthology.org/P19-1050.

[11] A. Saravanan, G. Perichetla, and K. S. Gayathri, "Facial emotion recognition using convolutional neural networks," *ArXiv*, vol. abs/1910.05602, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204509393.

[12] E. Pranav, S. Kamal, C. S. Chandran, and M. Supriya, "Facial emotion recognition using deep convolutional neural network," in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2020. DOI: 10.1109/icaccs48705.2020.9074302. [Online]. Available: https://doi.org/10.1109/icaccs48705.2020.9074302.

[13] D. Theckedath and R. R. Sedamkar, "Detecting affect states using vgg16, resnet50 and se-resnet50 networks," *SN Computer Science*, vol. 1, no. 2, Mar. 2020, ISSN: 2661-8907. DOI: 10.1007/s42979-020-0114-9. [Online]. Available: http://dx.doi.org/10.1007/s42979-020-0114-9.

[14] L. Zahara, P. Musa, E. Prasetyo Wibowo, I. Karim, and S. Bahri Musa, "The facial emotion recognition (fer-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (cnn) algorithm based raspberry pi," in *2020 Fifth International Conference on Informatics and Computing (ICIC)*, 2020, pp. 1–9. DOI: 10.1109/ICIC50835.2020.9288560.

[15] Y. Khaireddin and Z. Chen, *Facial emotion recognition: State of the art performance on fer2013*, 2021. eprint: arXiv:2105.03588.

[16] B. Koonce, "Resnet 50," in *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*. Berkeley, CA: Apress, 2021, pp. 63–72. DOI: 10.1007/978-1-4842-6168-2_6. [Online]. Available: https://doi.org/10.1007/978-1-4842-6168-2_6.

[17] A. Saroop, P. Ghugare, S. Mathamsetty, and V. Vasani, *Facial emotion recognition: A multi-task approach using deep learning*, 2021. eprint: arXiv:2110.15028.

[18] A. Vulpe-Grigoraşi and O. Grigore, "Convolutional neural network hyperparameters optimization for facial emotion recognition," in *2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, 2021, pp. 1–5. DOI: 10.1109/ATEE52255.2021.9425073.

[19] P. Bagane, S. Vishal, R. Raj, T. Ganorkar, and Riya, "Facial emotion detection using convolutional neural network," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 11, 2022. DOI: 10.14569/IJACSA.2022.0131118. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2022.0131118.

[20] S. P. Langute, *Facial emotion recognition using deep convolutional neural network*, Jan. 2022. [Online]. Available: https://norma.ncirl.ie/6214/1/sachinpralhadlangute.pdf.

[21] R. C. Lokesh Bejjagam, *Facial emotion recognition using convolutional neural network with multiclass classification and bayesian optimization for hyper parameter tuning*, Jun. 2022. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1677831/FULLTEXT02.

[22] C. Peng, Y. Liu, X. Yuan, and Q. Chen, "Research of image recognition method based on enhanced inception-resnet-v2," *Multimedia Tools and Applications*, vol. 81, no. 24, pp. 34 345–34 365, May 2022, ISSN: 1573-7721. DOI: 10.1007/s11042-022-12387-0. [Online]. Available: http://dx.doi.org/10.1007/s11042-022-12387-0.

[23] A. Ilham, S. Suryono, and C. E. Widodo, "A review of face expression recognition system recommendation for psychiatric disease," in *THE 6TH INTERNATIONAL CONFERENCE ON ENERGY, ENVIRONMENT, EPIDEMIOLOGY AND INFORMATION SYSTEM (ICENIS) 2021: Topic of Energy, Environment, Epidemiology, and Information System*, AIP Publishing, 2023. DOI: 10.1063/5.0125595. [Online]. Available: https://doi.org/10.1063/5.0125595.

[24] D. D. Patil, *Facial emotion recognition using deep learning*, May 2023. [Online]. Available: https://scholarworks.calstate.edu/downloads/b5645007n.

[25] H. M. Shahzad, S. M. Bhatti, A. Jaffar, S. Akram, M. Alhajlah, and A. Mahmood, "Hybrid facial emotion recognition using cnn-based features," *Applied Sciences*, vol. 13, no. 9, 2023, ISSN: 2076-3417. DOI: 10.3390/app13095572. [Online]. Available: https://www.mdpi.com/2076-3417/13/9/5572.

[26] A. Sultana, S. K. Dey, and M. A. Rahman, "Facial emotion recognition based on deep transfer learning approach," *Multimedia Tools and Applications*, May 2023. DOI: 10.1007/s11042-023-15570-z. [Online]. Available: https://doi.org/10.1007/s11042-023-15570-z.

[27] F. M. Talaat, "Real-time facial emotion recognition system among children with autism based on deep learning and iot," *Neural Computing and Applications*, vol. 35, no. 17, pp. 12 717–12 728, Mar. 2023, ISSN: 1433-3058. DOI: 10.1007/s00521-023-08372-9. [Online]. Available: http://dx.doi.org/10.1007/s00521-023-08372-9.

[28] H. F. T. Alsaadawı and R. Daş, "Multimodal emotion recognition using bi-lg-gcn for meld dataset," *Balkan Journal of Electrical and Computer Engineering*, vol. 12, no. 1, pp. 36–46, 2024. DOI: 10.17694/bajece.1372107.

[29] F. M. Talaat, Z. H. Ali, R. R. Mostafa, and N. El-Rashidy, "Real-time facial emotion recognition model based on kernel autoencoder and convolutional neural network for autism children," *Soft Computing*, Jan. 2024, ISSN: 1433-7479. DOI: 10.1007/s00500-023-09477-y. [Online]. Available: http://dx.doi.org/10.1007/s00500-023-09477-y.