Emotional Assessment in Children with Downs Syndrome Using Deep Learning During Dolphin Assisted Therapy

By

Faith Sanne Odhiambo

121058

Submitted in partial fulfillment of the requirements for the Degree of

Master of Science in Information Technology (MSc.IT) at Strathmore University

School of Computing and Engineering Sciences, Strathmore University, Nairobi. Kenya

June ,2025

Declaration and Approval

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Student's Name: **Faith Sanne Odhiambo**

Sign: Date: **27/05/2025**

Approval

The thesis of **Faith Sanne Odhiambo** was reviewed and approved by the following:

Dr. Joseph Onderi Orero,

Senior Lecturer, School of Computing and Engineering Sciences

Strathmore University

Eng. Dr. Julius Butime,

Dean, School of Computing and Engineering Sciences

Strathmore University

Prof. Bernard Shibwabo,

Director of Graduate Studies,

Strathmore University

Abstract

Understanding emotions is critical for supporting behavior and therapy in individuals with cognitive disabilities such as Down syndrome. However, accurately recognizing emotional states in this population remains a challenge due to subtle or atypical facial expressions. This study investigated automatic emotion recognition and assessment in children with Down syndrome undergoing Dolphin-Assisted Therapy (DAT). This work contributes to the development of personalized, data-driven tools to support therapeutic interventions in special needs populations. We proposed three approaches: An image-based approach that involved training raw images on a custom convoluted neural network, a feature-based approach that combined semantic emotion scores from DeepFace with geometric facial landmarks extracted via Mediapipe and combined approach of CNN embeddings and the features extracted. The CNN model trained on raw images only achieved 76% accuracy. The models trained on the feature set performed as follows; Random Forest: 62%, Support Vector Machine :59%, Multi-Layer Perceptron:50% and Dense Neural Networks:50%. Finally, the hybrid model (CNN with Dense fully connected layer) achieved the highest accuracy of 84%.

The results show that combining these complementary features improves classification performance, offering a more objective and interpretable method for assessing emotional engagement during therapy.

Table of Contents

List of Figures

List of Tables

List of Appendixes

List of Abbreviations

ASD	Autism Disorder
AU	Action Units
CHD	Congenital Heart Defects
CNN	Convolutional Neural Networks
DAT	Dolphin Assisted Therapy
DCNN	Deep Convolutional Neural Networks
DS	Down Syndrome
ECG	Electrocardiogram
EEG	Electroencephalogram
ER	Emotion Recognition
ERS	Emotion Recognition System
FACs	Facial Action Units
LSTM	Long-Short Term Memory
MFCC (MFCCs	Mel-frequency cepstral coefficients
MLP	Multi-Layer Perceptron
SVM	Support Vector Machine
PAD	Pleasure-Arousal-Dominance
PCA	Principle Component Analysis
RF	Random Forest
WHO	World Health Organization

Acknowledgements

I extend my deepest gratitude to my supervisors, Dr. Joseph Onderi Orero and Professor Ismael Ateya, for their invaluable guidance and unwavering support throughout this thesis journey. Their expertise and mentorship have been instrumental in shaping my research.

I also wish to express my sincere appreciation to Dr. Jesus Moreno Escobar and Instituto Politécnico Nacional (The National Polytechnic Institute) for their generous provision of data. Your invaluable advice, guidance, and patience have significantly contributed to this work.

I am profoundly grateful to Strathmore University for providing essential resources and fostering an environment that has enabled me to grow academically and establish myself as a researcher. Finally, my heartfelt appreciation goes to the Master of Science in Information Technology Class of 2025 with special mentions : Sandra Gathoni, Salome Chemiat and Noela Ahindukha for their unwavering support and encouragement throughout the thesis writing process.

Chapter 1 Introduction

1.1 Background

Recognizing and interpreting emotions is fundamental to human interaction and experience, yet it is particularly challenging to a select few who have communication barriers, such individuals with Down syndrome. Cognitively and mentally atypical persons can often demonstrate a unique emotional expression profile, which can make it hard to interpret. The main reason is that traditional emotion recognition methods primarily rely on verbal cues or typical visual cues. As society pushes for more inclusive and adaptive systems, emotion detection and recognition are becoming increasingly important across different disciplines such as human-computer interaction (HCI), adaptive learning systems, mental health monitoring and clinical therapy.

Within therapy, emotional assessment remains crucial especially for children with Down syndrome (DS), as it not only helps clinicians and caregivers understand the child's emotional state but also significantly contributes to the overall well-being. Some common emotional assessment methods are facial expression analysis, physiological signal monitoring, and behavioral observation. (Antonarakis et al,2019).

Several therapeutic approaches are used to support children with DS, including speech therapy, occupational therapy, and music therapy. One of the new therapies: Dolphin-Assisted Therapy (DAT), has gained attention for its potential benefits in improving emotional engagement and cognitive function. In fact, Zambrano (2018), asserted that interactions with dolphins can enhance social skills, reduce anxiety, and stimulate sensory processing in children with developmental disorders. Dolphin Sound Therapy on the other hand has been explored for its calming effects on children with DS. While promising, further

research is necessary to establish the long-term efficacy of these therapies in improving emotional and cognitive outcomes in children with DS.

The past few decades have seen a steady increase in the development of algorithms capable of detecting and analyzing emotions through a range of data sources such as facial expressions, speech, body language (gesture) and physiological signals such as ECG and EEG. (Mahalakshmi, 2017). At the moment, facial expressions, which is visual, is the most used modality in emotion detection systems. How they work is by utilizing deep learning models to map the position or change in facial muscles to a particular emotion.

Instead of relying on a single mode of detection, there are new approaches that consider the complexities and variability of human emotions. Oviatt's Multimodal Interaction Theory laid the groundwork for this shift by emphasizing the importance of combining different sensory inputs. Despite advances in emotion detection there is still a gap in the development of such algorithms that are tailored for individuals with down syndrome. (Oviatt & Cohen, 2015)

Following the appreciation of the complexities of human emotion, machine learning research has moved from simplistic models to more complex, hybrid approaches. Deep learning frameworks have been used for both unimodal and multimodal approaches, particularly neural networks. Additionally, significant advancements have been made in the development of datasets for emotion detection and classification. (Paredes, Caicedo-Bravo, & Bacca, 2023). Early datasets primarily focused on single modalities such as facial expressions (e.g., CK+ and FER2013), speech signals (e.g., RAVDESS and EMO-DB), and physiological signals like electroencephalography (EEG) (e.g., DEAP and SEED). (Zhang, Zhang, & Li, 2018).

The use of neural networks in emotion detection has really expanded our understanding of how emotions work in atypical individuals. For example, one study in 2015 used a multi-modal approach and explored the relationship between physiological arousal, facial expressions, and self-awareness in emotional experiences of children with autism spectrum disorder (Jain et al., 2015). However, emotion detection comes with its own set of challenges. Modeling complex emotions is difficult due to the very nature of emotions. Human emotions are dynamic; they overlap and are heavily influenced by individual differences meaning they will vary from one person to the next. Another challenge lies in limited data for certain demographics, such as children with Down , which restricts current model's generalization. Lastly, how to effectively combine the different modalities is still a challenge due to variations in the signals and synchronization issues. More research is needed to address the gaps.

1.2 Problem Statement

Children with Down Syndrome often face difficulties in expressing and regulating emotions due to their unique emotional barriers, which can lead to frustration and behavioral issues. There is therefore a great need to develop more comprehensive tools to assist caregivers such as clinical therapists in identifying emotional cues for effective support and care. (Esbensen et al., 2024). This is the main gap being addressed in this study.

Most (current) emotion detection systems rely on facial expressions and vocal cues, which may not be fully dependable for children with Down syndrome due to their atypical emotional expressions and communication delays. Furthermore, existing research and technologies do not adequately address the emotional variability and complexities in this population, leaving a gap in understanding their real-time emotional responses during therapy. (Moreno Escobar et al., 2023; Paredes et al., 2022).

1.3 Objectives

1.3.1 General Objective

To develop an emotion assessment system using deep learning for children with Down Syndrome to that can accurately detect and interpret emotional states during Dolphin Assisted Therapy

1.3.2 Specific Objectives

i. To review existing emotion recognition and assessment approaches and their implementations.

ii. To design an emotion assessment model considering different modalities for children with Down Syndrome

iii. To develop the designed emotion assessment model using deep learning for children with Down syndrome

iv. To validate the developed model

1.4Research Questions

i. What are the existing emotion recognition and assessment approaches and implementations?

ii. How can an emotion assessment model considering the identified modalities for children with Down Syndrome be designed?

iii. How can the designed emotion assessment model for children with DS be developed using deep learning?

iv. How can the developed model be validated?

1.5Justification

Paredes et al. (2023) asserts that effective therapy relies on understanding the emotional state of a patient which in turn assists the therapist to connect and adapt their

approach. For children with Down Syndrome, who often exhibit atypical emotional expressions or delays in emotional responses, a deep learning model can more accurately capture their emotional state which can be useful for tailoring therapeutic interventions. For individuals with various disabilities, control, regulation, knowledge, and expression of emotions can improve their quality of life.

Such an algorithm should equip therapists with valuable insights into the emotional state of children with DS, enabling them to develop effective strategies to facilitate the completion of assigned activities. Additionally, this emotional assessment throughout therapy can aid in identifying patterns or triggers that may impact the child's progress, allowing for timely adjustments in the treatment plan which can gown a more supportive and empathetic therapeutic environment (Rodrigo-Claverol et al., 2020).

1.6 Scope and Limitations

This study only focused on the development of an emotion assessment model for children with Down Syndrome during Dolphin Assisted Therapy. The study did not cover vocal cues, text-based and physiological modalities in detail due to resource limitations; specifically, limited availability of such data. This study considered both unimodal (single-source) and multimodal (multiple-source) approaches and compared them to determine the optimal approach. As this research was centered around Dolphin Assisted therapy / Dolphin Sound Therapy as the testing environment, the model's generalizability in other therapeutic environments may require further validation and adaptation.

Chapter 2 Literature Review

2.1 Introduction

Emotion recognition and assessment involves analyzing behavioral and physiological signals to classify emotional states. Research in this field has evolved significantly, with early studies demonstrating the effectiveness of using facial expressions, speech patterns, and physiological signals for emotion detection. Advances in machine learning have further enabled the development of models that analyze facial expressions, vocal tone, and neurophysiological signals, offering deeper insights into emotional responses in children with Down syndrome.

2.2 Theoretical Literature Review

This section reviewed the relevant theories that form the foundations of design and development of emotion recognition systems with a focus on facial emotion modelling. It also explored how these theoretical perspectives can be adapted for children with Down Syndrome to bridge the emotional and communication gaps that exist.

2.2.1 Theories of Emotion

Psychologists define emotions as a state of belief which manifest as a change in psychology which in turn causes a physical change. Human emotion is a conscious appraisal which involves physiological arousal and expressive behavior. (Lerner et al., 2015) The main theories of emotion can be grouped into three main categories.: Physiological, Neurological and Cognitive. The Physiological category consists of the James-Lange and Cannon-Bard Theory and proposes that emotions are a function of what happens within. Neurological theories such as Facial Feedback and Cognitive appraisal theories argue that the brain is responsible. Finally, the last category brings in the relationship between

cognition and emotions. These theories can further be grouped under two models for analyzing emotions; Categorical Models and Dimensional Models. (Mahalakshmi, 2017)

2.2.1.1 Categorical Emotion Models

Categorical Emotion Models depict emotions as discrete states such that, at a particular point in time, the emotion which is selected is that which best represents the feelings being conveyed. This model makes use of labelled emotions according to the Basic Emotion theory which proposes six primary emotions: happiness, sadness, anger, fear, surprise and disgust. (Ekman et al., 1969). While most of the research in the field of Affective computing uses the above set of emotions there are some that have moved to more 'domain-specific' labels. An example is the replacement of fear and disgust with boredom or irritation in the learning and gaming domain as it better suits the application context. The biggest drawback of such models is that the boundaries that come with classification oversimplify the complexity of human emotions. Due to the dataset and the simplicity of these models, this is the model that the developed solution is based on.

2.2.1.2 Dimensional Emotion Models

Dimensional models link various effects in a common dimensional space. It may be two dimensional; arousal and valence or three dimensional: arousal, valence and power. Plutchik presents a model that uses the activation-evaluation dimension which informs the emotion wheel. The Mehrabian model uses PAD; Pleasures -Arousal-Dominance. Here the dominance domain is like the valence in Russell's model and assesses the level of control a person has with regard to the state they are in. Although this approach does not associate an affect with a labelled emotion, it has the advantage of capturing finer emotion concepts and therefore covers a wider range of emotional states. It also shows the relationship of various

emotional states within that dimensional space. For example, the figure below shows the distribution of emotions within the Arousal-Valence Domain. (Mehrabian & Russell,1974)

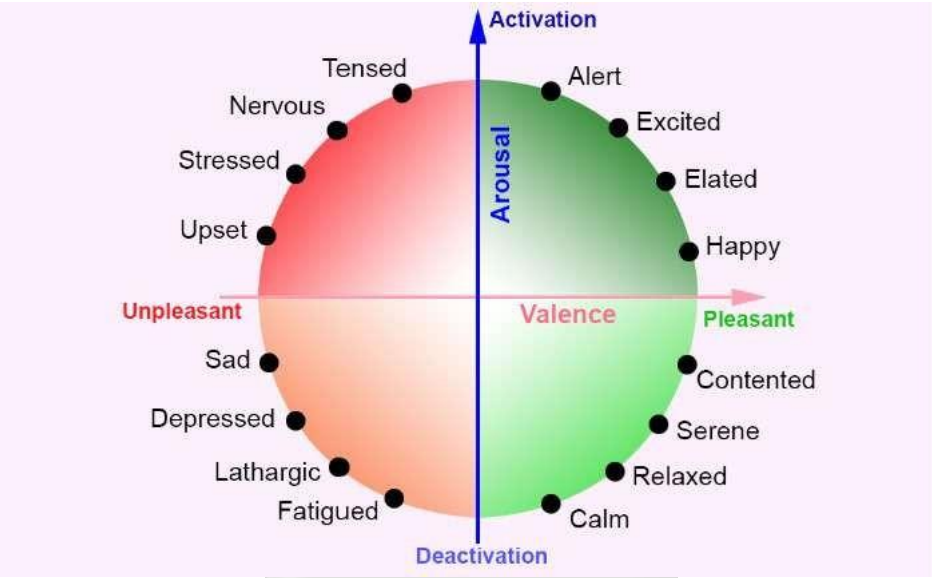


Figure 2.1. Distribution of Emotions around the 2D Arousal-Valence (Mahalakshmi, 2017)

This study preferred to go with the Categorical model which has its roots in the Basic Emotion Theory proposed by Ekman because of the simplicity of discrete emotion which facilitated easier classification and consequent comparison between different emotional states.

2.2.2 Facial Emotion Modelling and Recognition

Emotion recognition refers to the process of identifying and interpreting human emotions through various modalities, such as facial expressions, gestures, voice, text, body language, and physiological signals through EEG and ECG. Building on foundational theories like Ekman's basic emotion theories, traditional systems have relied heavily on facial expressions to classify emotions. These systems use computer vision algorithms to detect facial landmarks and classify expressions into predefined emotional categories. (Ekman, 1969)

The last few years have seen an increase in the application of emotion recognition in the fields of medicine (particularly psychology and clinical therapy) and affective computing. This can be attributed to research that supports the growing importance of accurately identifying emotions to understand the influence they have on diseases and vice versa. An example is that in psychology, the manic episode exhibited by a patient with bipolar disorder are being examined as extreme expressions of basic emotions such as happiness and sadness.

For facial emotional modelling, emotion recognition involves two crucial steps: feature extraction and classification. Feature Extraction involves the identification of features or attributes specific to a given emotional expression while classification maps the features into the labelled categories or classes of emotional expression. (Guo,2024)

2.2.2.1 Modelling Facial Expressions

Facial expression recognition uses various computer vision techniques to analyze key facial landmarks and map them to specific emotions. One popular method for this is the Facial Action Coding System (FACS), which categorizes facial muscle movements into Action Units (AUs). As developed by Paul Ekman and Wallace V. Friesen in 1978, FACS is a anatomically based system for describing all visible facial movements. As a way of standardizing facial expressions across different people , It breaks down facial expressions into individual muscle movements.

FACS has already been widely used in research and as the basis for real applications. For example, Liu et al. (2020) reviewes how FACS has been applied to study human emotional reactions to consumer products. Additionally, Breuer and Kimmel (2017) explored how deep learning models interact with FACS.

Lien et al. (1998) was one of the earliest works to propose a method that recognizes facial expressions by identifying these Action Units, by analyzing both upper and lower facial movements. However, despite the progress made in automating this coding system, there are still challenges when it comes to accuracy and reliability, For instance , Cohn et al. (2019) discussed the potential pitfalls of automated FACS coding stating factors such as facial expression variability, subtle and complex movements and data limitation. The paper further highlighted the need for thorough validation to avoid misinterpretations in emotional research.

An older technique is Principal Component Analysis (PCA). It was introduced by Karl Pearson in 1901. This statistical technique works by reducing the dimensionality of facial images without compromising essential features. Each facial image is converted into a high-dimensional vector which undergoes a series of function to extract the principal(main/essential) components. The upside is that instead of storing the full images, only the Eigenface values are stored.

In their 2001 study, "A Principal Component Analysis of Facial Expressions," Calder et al. examined facial expressions from the Ekman and Friesen (1976) collection using PCA to understand the structure behind facial expressions. Their findings showed that PCA could support facial expression recognition effectively. The conclusion revealed that the components used for coding facial expressions are different from those used for facial

identity. These techniques cannot be used without adaptation to certain populations; as to improve the overall accuracy of emotion recognition systems.

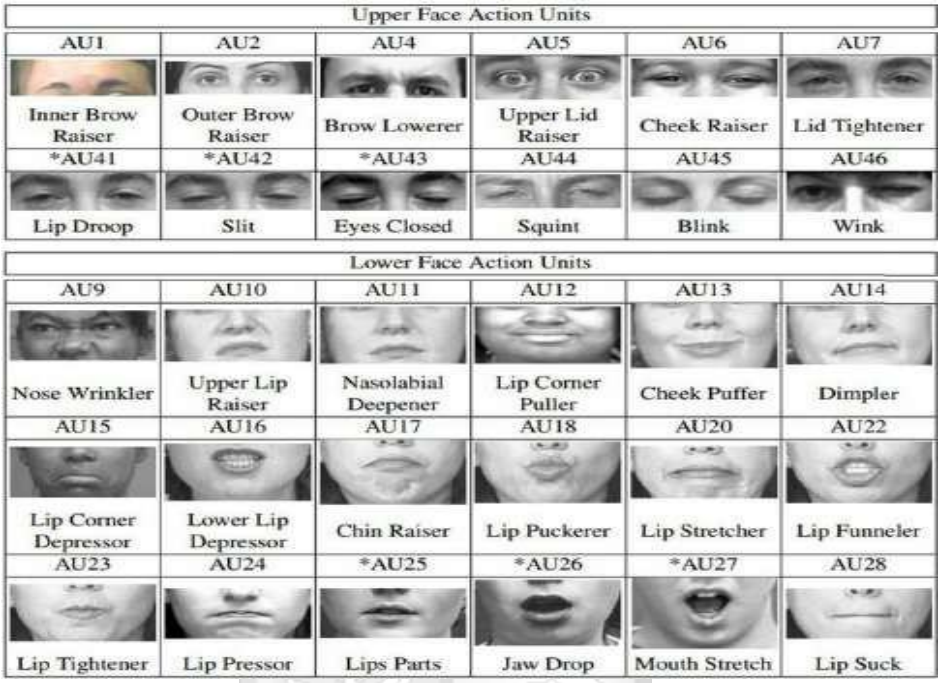


Figure 2.2 Facial Action Units AU1 to AU28(Jacintha et al., 2019)

This showed that the PCA system's recognition rates for individual expressions were comparable to human rates, supporting the effectiveness of PCA in modeling facial expression recognition. (Calder et al., 2001).

2.2.2.2 Modelling Multiple Modalities

Multimodal Emotion Recognition combines data from multiple modalities such as facial expressions, physiological signals, speech, body movements to identify the correct emotional state. There is an increase in the development of emotion recognition systems that incorporate various forms of data, such as the ones discussed in the previous section in a bid to create systems that are more comprehensive and optimized (J. Pan et al., 2024; Vazquez-Rodriguez, 2021; Yang et al., n.d.) .

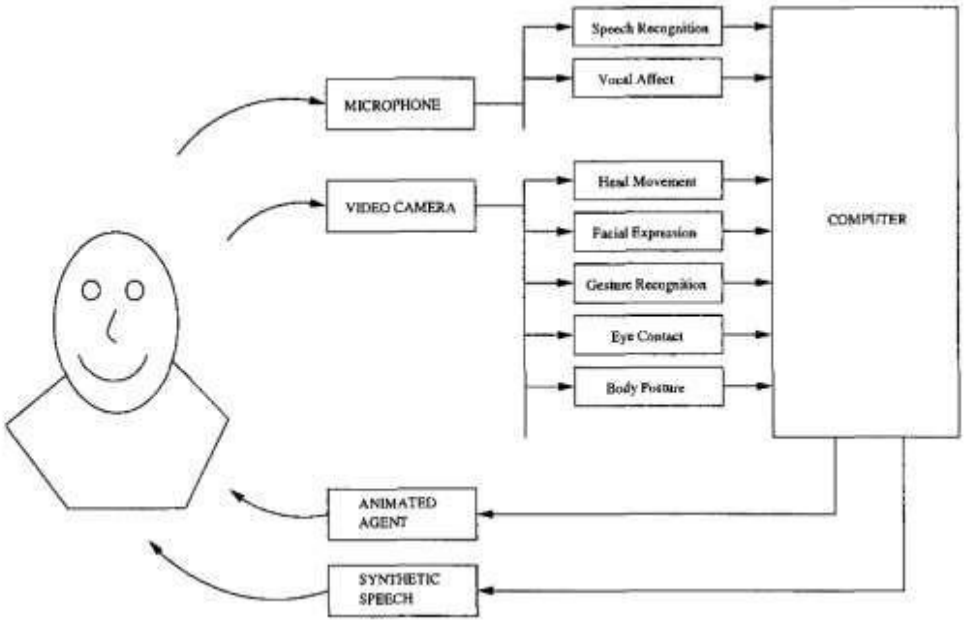


Figure 2.3 Simple multimodal framework for Multimodal Emotion Detection

A key challenge when dealing with multiple modalities lies in identifying the weight a particular modality carries when providing information about a certain emotion. Another typical issue is that multimodal data is usually processed separately. According to, for proper multimodal analysis, the signals must be considered mutually dependent and must therefore be combined in a contextual manner at the end. When combining multimodal data, inter-modality, cross-modality and missing data should be taken into consideration. (Pawłowski et al., 2023)

2.2.3 Fusion Techniques

Fusion techniques are essential for integrating information from multiple modalities in emotion recognition systems**.** There are two primary fusion techniques: early and late fusion.

2.2.3.1. Early Fusion

In early fusion, also known as low level fusion, raw data (in form of vectors) are combined at the feature extraction stage before classification. This method captures the interdependence between modalities at an early stage. The advantage of early fusion is that data processing is more comprehensive. The challenge associated with this technique is that it can lead to high-dimensional datasets which make selecting appropriate features difficult. (Pawłowski et al., 2023).

2.2.3.2. Late Fusion

Late fusion, also known as high level fusion, involves the processing of each modality independently and then combining the outputs at the decision-making stage. Compared to early fusion, this method is less demanding (computationally), but the drawback is that it may miss cross-modal interactions that could be significant. Late fusion is modular and therefore fault tolerant; if a modality fails, the system can still rely on another modality. Additionally, different models can be tailored for each modality allowing for specialized processing techniques for each data type. (Zhang.et al, 2020)

2.2.3.3 Hybrid Fusion

As the name suggests, this technique combines both aspects of early and late fusion techniques. The data from various sources is first processed to extract relevant features. These features can be concatenated and then fed into separate models for further analysis. The outputs of these models can then be combined to make final predictions, allowing for

both the exploitation of inter-modal relationships and the independent strengths of each modality. Hybrid techniques have been applied to medical imaging where MRI, CT, and PET provide complimentary information. The challenge is in the implementation and the computational demand, especially in real time applications. (Sahu & Vechtomova, 2019)

2.3 Empirical Literature Review

The empirical literature review provides an analysis of existing research in emotion recognition, particularly in neurodiverse populations such as children with Down syndrome. In addition to exploring various methodologies, algorithms, and feature extraction techniques, it also covers the unique emotional expression patterns of individuals with Down syndrome.

2.3.1 Overview Down syndrome.

Down Syndrome or Trisomy 21 is a genetic condition characterized by the presence of an abnormal number of chromosomes in a cell. Instead of the typical diploid number of chromosomes (46 in humans, arranged in twenty-three pairs), they have aneuploid cells with an extra chromosome present. (Boato et al., 2022).

There are several hypotheses surrounding the cause of this occurrence. The gene dosage imbalance hypothesis states that it is due to an increased number of genes in chromosomes which disrupts development. There is also the critical region hypothesis which highlights specific chromosomal areas, particularly 21q21.22, which are associated with various clinical features of Down syndrome. Research points out that both the critical region hypothesis highlights specific chromosomal areas, particularly 21q21.22, which are associated with various clinical features of Down syndrome. Research shows that both multiple genes and multiple critical regions can contribute to the phenotypical traits of the disorder. (Antonarakis et al., 2020; Esbensen et al., 2024)

2.3.1.1 Clinical Features

Several clinical conditions are associated with DS as it affects different systems of the human body ranging from intellectual, neurological, congenital, gastrointestinal, and facial features, among others.. The figure below shows some of the characteristic facial and physical features of children with DS.

Figure 2.4 A drawing of the facial features of Down Syndrome contributed by the Centers for Disease Control and Prevention, National Center on Birth Defects and Developmental Disabilities (Public Domain) (Akhtar & Bokhari, 2023)

Congenital Heart Defects (CHD) are the most prevalent; 50% of babies born with DS have CHD. For neurological disorders, they are present with reduced brain volume which affects motor development, joint laxity and muscle development. Almost all patients with DS have difficulties in learning and memory.

2.3.1.2 Treatment and Management of DS

There is no known cure for DS, treatment is based on presented symptoms. A multidisciplinary approach is taken to manage the conditions that these individuals have. This includes karyotyping, parental/guardian education and awareness, pediatric, gastroenterologist, neurologist, and psychiatrist care. Children with DS benefit from

occupational, physical, speech and behavioral therapy as well as special education. Recent advances in healthcare and technology have significantly increased the rate of survival for cases of DS but they still have a shorter life expectancy compared to healthy individuals. (Antonarakis et al., 2020; Esbensen et al., 2024)

2.3.1.3 Social and Emotional Development

Children with DS often face social challenges stemming from their cognitive and developmental deficiencies. Research supports the idea that they may fail to grasp social cues, language and emotions which affect their relationships. Together with behavioral and sensory challenges may make navigating social interactions an issue. (Sideropoulos et al., 2023) A cross-sectional study involving 6-year-olds that compared the social capabilities of healthy children and those with DS found that the latter were generally deficient. The study found that they had more social problems attributed to their language impairment. Therefore, they are often observed to rely on non-verbal communication such as facial expressions or gestures to communicate as their language and speech still develops. (Day et al., 2022)

2.3.2 Applications for Facial Emotion Recognition

Emotion recognition plays a crucial role in affective computing, enabling machines to interpret and respond to human emotions. FER can be spatial or spatiotemporal. The former uses a single frame to extract the AUs while spatiotemporal captures the facial actions within a temporal domain to capture the dynamic features. Certain studies have exploited dynamic features which use motion pictures from history; used a combination of appearance and motion from local information and developed Harr features which are dynamic. Most of the studies that have been done have utilized a whole video sequence containing a single expression.

For instance, Zhang et al. (2019) proposed a spatiotemporal approach for FER using a combination of facial landmark tracking and temporal feature extraction techniques. The study showed the movement of facial landmarks across frames significantly enhanced emotion classification performance in emotions like surprise and fear. This greatly outperformed static models by capturing the motion dynamics in facial expressions.

Nguyen et al. (2020) also investigated the effectiveness of combining spatial and temporal features in FER, using deep learning methods. They developed a hybrid architecture that combined both Convolutional Neural Networks (for spatial feature extraction) and LSTM (Long Short-Term Memory) networks for temporal feature extraction, using video data from the well-known AFEW dataset. The hybrid approach also outperformed purely spatial methods in detecting emotions such as happiness and sadness, especially in much longer video sequences, where temporal context played a more significant role .

Spatial FER has evolved due to such advanced approaches and the different datasets that have been curated such as the Cohn-Kanade (CK), AFEW and Facial Expression Recognition 2013, FACES (Agung et al., 2024; Guo et al., 2024; Jacintha et al., 2019). As explored in 2.2.2.1, facial expression recognition analyzes key facial landmarks to correlate them to a specific emotion.

2.4 Review of Architectures of Emotion Recognition Systems

Recognition systems, also known as pattern recognition systems, are designed to classify and identify features or patterns in data. These systems have been widely used in various fields, such as handwriting recognition, face recognition, speech recognition, and computer vision. This study proposes following a similar approach for emotion recognition with the basic architectural elements of a face recognition system which include face

detection, preprocessing, feature extraction, and face recognition (Koutroumbas & Theodoridis, 2008; Rao & Reddy, 2011; Kasar, Bhattacharyya & Kim, 2016).

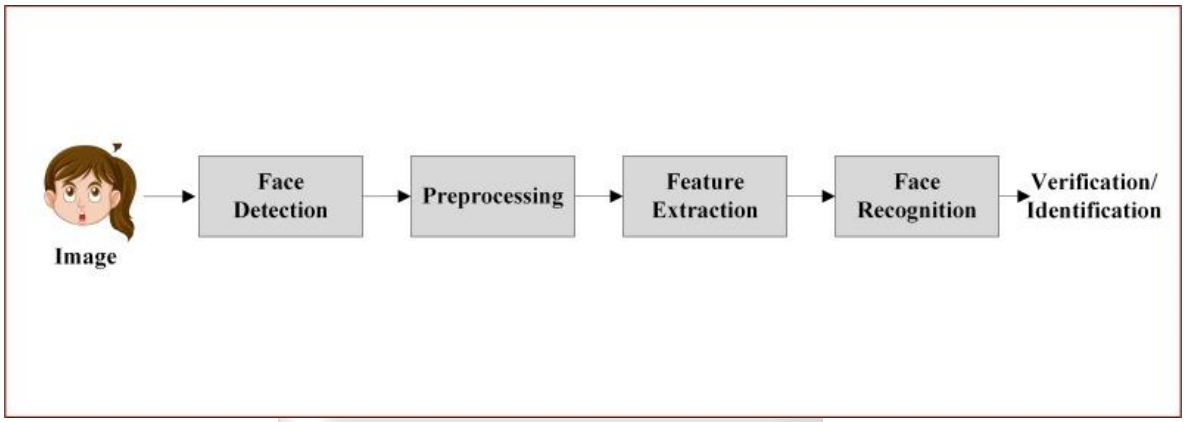


Figure 2.5 Basic Architectural Structure of Recognition systems

This architecture is suitable for the development of a recognition system that utilizes only one source or modality. In the case of other modalities, a study in 2023 compared techniques for multimodal data fusion and presented architectures for unimodal, bimodal and trimodal systems. The study utilized the Amazon Reviews dataset. They noted that the only difference in architecture is that unimodal models do not have a model on top of their outputs in the late fusion approach. (Pawłowski et al., 2023).

The study highlighted that the bimodal and trimodal models outperform unimodal models in every category (accuracy metric). The difference in performance was significant per class where unimodal models yield close results. However, they found that using trimodal approach led to little or significant change.

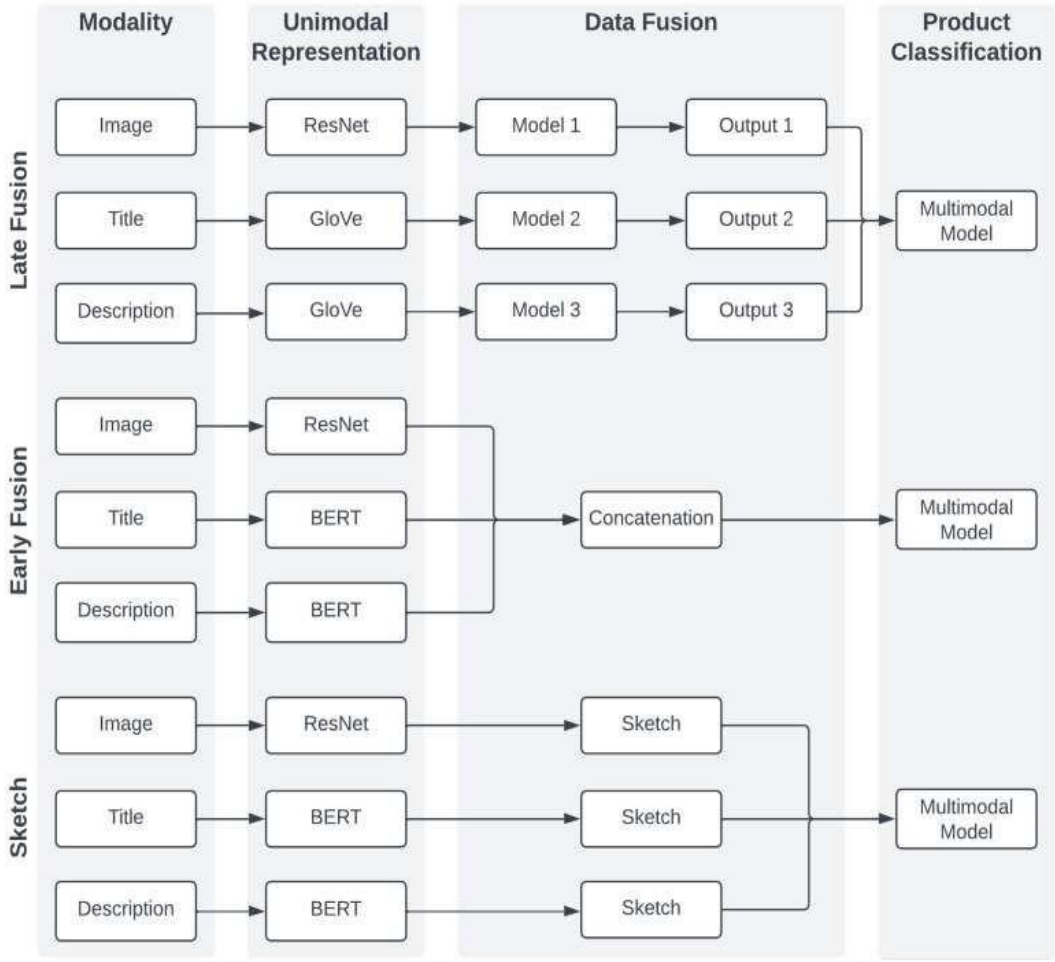


Figure 2.6 Comparative architecture for Early and Late Fusion Techniques in Multimodal Systems *(Pawłowski et al., 2023)*

2.5 Review of Algorithms used for Emotion Recognition Systems

A review of Algorithms used over the past years indicated that most systems used CNNs or Modified CNNs. (Islam et al., 2021a; Ouzar et al., n.d.; J. Pan et al., 2024; Verma et al., n.d.; Yang et al., n.d.).

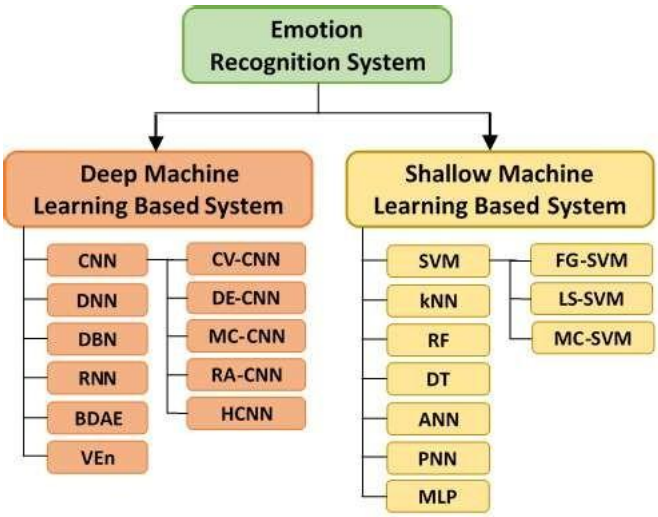


Figure 2.7 Categories of general emotion recognition systems (Islam et al., 2021a)

2.6 Deep learning Algorithms

This entails a subset of machine learning that use artificial neural networks, particularly multi-layered ones, to learn complex patterns and features from data, automatically, without explicit human-driven feature engineering.

Neural networks, computational models inspired by the way the human brain processes information, are used for emotion recognition by learning complex patterns in facial expressions. It consists of layers of interconnected nodes (neurons), which process data and learn patterns. Neural networks are the foundation of deep learning and are used for tasks like image recognition and language processing. CNNs are specialized for processing visual data like facial images, widely used for facial expression recognition in emotion detection systems. RNNs, including LSTMs, are well-suited for modeling temporal dependencies. (Bishop,2006)

2.6.1 Convolutional Neural Networks (CNNs):

Convolutional Neural Networks (CNNs) have become a cornerstone for processing images and topographical maps. This type of neural network is designed to process spatial data such as images topographic maps. Their ability to apply filters that extract localized features allows these networks to excel in various analytical tasks. The fundamental operation of a CNN is the convolution operation, which can be mathematically expressed as follows (Bishop,2006)

$$y_{i,j} = (f * x)_{i,j} = \sum_m \sum_n f_{m,n} \cdot x_i + m, j + n \quad (1)$$

In this equation, f denotes the filter or kernel, x symbolizes the input signal (in this case, images in the case of visual data), and y represents the output feature map. This feature extraction approach enables CNNs to perform a sort of hierarchical learning, which is helpful in identifying spatial patterns within the raw data. Usually, after several convolutional layers, the output is channeled into a fully connected or dense layer (s), which can be expressed mathematically as follows:

$$y = \sigma(W \cdot x + b) \quad (2)$$

In this equation, W represents the weight matrix, x is the input vector, b denotes the bias term, and σ is the activation function, often selected as ReLU or SoftMax. The advantage of CNNs over other algorithms is their capacity to learn spatial dependencies ***directly*** from the raw signals, which minimizes or can entirely eliminates the need for manual feature extraction.

There is research that supports that CNNs can enhance emotion classification accuracy by up to 20% in comparison to other more traditional methods such as Support Vector Machines (SVMs). (Liu et al., 2020).

2.6.2 Recurrent Neural Networks (RNNs):

Recurrent Neural Networks (RNNs) is a type of neural network that was designed to process sequential data. What makes it stand out is the ability to preserve temporal dependencies through the introduction of hidden states. The recurrence relation that characterizes an RNN can be expressed as:

$$h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot x_t + b) \quad \textbf{(3)}$$

Where?

h_{t-1} represents the previous hidden state,

x_t is the input at time step t ,

W_h and W_x are the corresponding weight matrices,

b reflects the bias term,

σ is the activation function, frequently using tanh or ReLU.

Apart from emotion recognition, these RNN's have been used for ERP analysis.

There is research that shows that RNN-based models achieve classification accuracy improvements of 15–25% over traditional feedforward networks when analyzing data (Karim et al., 2020). However, a major drawback is that traditional RNNs struggle with long-term dependencies due to an issue known as the vanishing gradient problem. This is what led to the development of variants as discussed below.

2.6.3 Long Short-Term Memory (LSTM) Networks

Long Short-Term Memory (LSTM) networks are a specialized variant of RNNs designed to address the vanishing gradient problem and capture long-term dependencies in sequential datasets. LSTMs introduce cell states and gated mechanisms, defined by:

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big) \ (forget\ gate) \quad \textbf{(4)}$$

$$\cdot\, i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \ (input\ gate) \quad \textbf{(5)}$$

$$C_t = f_t \cdot C_{t-1} + i_t \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \ \ (cell\ state\ update) \quad \textbf{(6)}$$

These kinds of neural networks are well for classification tasks such as cognitive load estimation and neurological disorder detection. They have demonstrated accurate improvements of 10-15 percent over standard RNNs. (Bishop,2006)

2.7 Shallow Learning Algorithms

Shallow learning algorithms, also known as traditional machine learning, are models with a single or few layers of transformation, focusing on extracting patterns from input data without extensive layer-by-layer analysis.

2.7.1 Multi-layer Perceptron

An MLP is a type of feedforward artificial neural network with at least three layers: an input layer, one or more hidden layers, and an output layer. MLPs are fully connected networks where each neuron in a layer is connected to every neuron in the next layer. MLPs are useful for classification tasks where the relationship between input features and target labels is non-linear.

As data flows from input to output, each hidden layer performs a weighted sum of its inputs followed by an activation function such as SoftMax, ReLU or Sigmoid. SoftMax is used in the output layer for multi-class classification tasks, ensuring the outputs sum to 1 and can be interpreted as probabilities. MLP learns through backpropagation, computes the gradient of the loss function with respect to each weight by the chain rule and adjusts the

weights accordingly to minimize the loss. As it is also a classical neural network, the output of the final layer is mathematically calculated using equation *(2)*

While MLPs are often used as shallow models in simpler tasks, their performance improves significantly with deeper architecture and larger datasets. Bishop,2006)

2.7.2 Support Vector Machines

SVMs are supervised learning models used for classification tasks. They work by finding the hyperplane that best separates data points of different classes in a high-dimensional space. SVMs are particularly useful in classification tasks with relatively small datasets and high-dimensional input data. They have been effectively applied in facial expression recognition tasks (Jain et al., 2019), where they offer reliable performance for tasks with clear decision boundaries. SVM is most commonly used and effective because of the use of the Kernel method which basically helps in solving the non-linearity of the equation in an extremely easy manner. The most common kernel functions are the linear kernel, polynomial kernel, and RBF (Radial Basis Function) kernel.

SVM uses an optimization function to maximize the distance between the hyperplane and closest data point while reducing errors in classification.

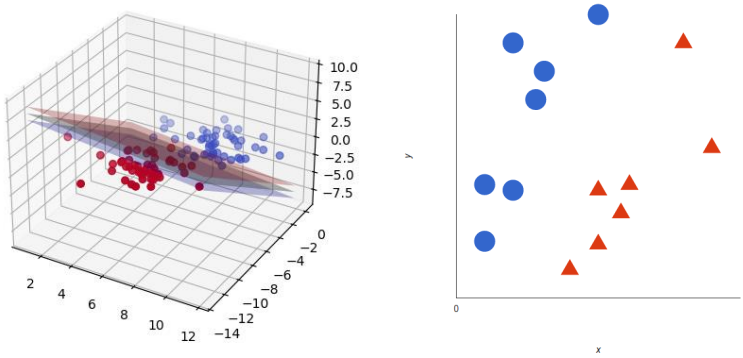


Figure 2.8 SVM Hyperplane in a 3D space vs 2D space.

The mathematical function is as follows:

$$y_{i,j} = min\frac{1}{2} \mid\mid w \mid\mid^2 subject\ to\ y_i(w \cdot x_i + b) \geq 1\ \forall_i \quad \textbf{(7)}$$

Where:

w is the weight vector,

x_i is the input feature vector,

y_i is the target label (+1 or −1),

b is the bias term.

2.7.3 Random Forest

Random Forests are an ensemble learning method that uses multiple decision trees to improve classification accuracy. A decision tree makes predictions by recursively splitting the data based on the best split which is essentially the feature that results in the best separation. The algorithm uses a given criteria, such as the Gini impurity index or a cross-entropy loss function. Each tree is trained on *a random* subset of the data, and their predictions are aggregated for a final output. This method has been found to be very effective for tasks where tabulated data may have more complex relationships or high-dimensional features. RFs are less prone to overfitting compared to individual decision trees, making them suitable for smaller datasets with varied features as stated earlier. They are often used as a in combination with facial action units or facial landmarks (Huang et al., 2020).

The equation below calculates the Gini impurity.

$$Gini(S) = 1 - \sum_{i=k}^{k} p_i^2 \cdot \quad \textbf{(8)}$$

Where p_i is the probability of class i in the set.

2.8 Challenges in Existing Methods and Techniques

The existing techniques still struggle with capturing the full complexity of human expression. For Feature extraction, PCA, does a well and reduces the dimensionality in given data but its strict structure means that it often overlooks non-linear facial variations. Many researchers have asked the question of whether alternative methods like Independent Component Analysis (ICA) could offer a better way to untangle facial features, though research in this area is still lacking (Liu et al., 2020). Another potential but challenging avenue is the combination of PCA with deep learning models like LSTMs (which can capture long term dependencies). This is very promising but so far, most of these studies have been tested on standard datasets, not on individuals with atypical facial muscle movements(Karim et al., 2020).

When it comes to algorithms, the common workaround has been to combine CNNs with LSTMs, but the issue of typical datasets still arises with very little generalizability to other scenarios. (Yousefi et al., 2019). Meanwhile, traditional Hidden Markov Models (HMMs) have been used to track facial action units (AUs) over time, but they usually model each AU independently, ignoring the fact that facial expressions involve coordinated muscle movements. There is still a general gap in developing models that can capture these interdependencies.

CNNs are very good in identifying spatial patterns in brain activity, but they do not do well with sequential data as we have seen in the section on algorithms. They have limits when it comes to their ability to model how emotions evolve over time (Karim et al., 2020). Researchers have tried blending CNNs with RNNs to fix this, but so far, these models have been heavily dependent on the dataset on which they were trained. Without a standardized approach to optimizing these architectures, it is hard to tell whether their performance is

truly generalizable. Attention mechanisms could help improve the way these models process long-term dependencies, but this area is still underexplored.

There is the bigger challenge of making different modalities work together effectively in a multimodal system. While deep learning has provided some fusion techniques, the way these modalities interact is still not well understood because the is no linear relationship between complex modalities (Nijland et al., 2020). On top of that, most of these systems do not consider interpersonal differences in how emotions are expressed, which makes it difficult to apply them to populations with unique characteristics. This is partly due to the ethical and logistical challenges of collecting data from such groups, but even if the data were available, there is still no universal method for designing multimodal emotion recognition models that can handle diverse datasets (Yousefi et al., 2019).

2.9Related works

2.9.1 Healthcare

In healthcare, non-invasive emotion recognition system are being integrated as part of new treatments or therapy especially for the management of cognitive disabilities and mental illness.

One study focused on using Deep Convolutional Neural Networks (DCNN) to analyze facial expressions of children with Down Syndrome during DAT demonstrating that ML models could effectively recognize emotional states from facial cues. The study uses Deep CNN to predict emotions and achieved an accuracy of 72%. This study also integrated the use of EEG signals which increased accuracy by 9 percent. The output provides invaluable insight for therapists and caregivers that guide treatment/management plans. (Moreno Escobar et al., 2023).

Mental illness is a global health concern that has devastating effects on the physical, emotional and social well-being of an individual. Emotion recognition systems have been used for detection of depression and also anger detection in patients. Another study proposed emotion recognition system for profound intellectual and multiple disabilities using motion and heartbeat interval (R-R interval). They used a random forest classifier and low-resource sensors. The findings indicated that the concept of emotion recognition systems is beneficial when communication with profound intellectual and multiple disabilities. (Tanabe et al., 2024)

2.9.2 Education

Emotion recognition systems are increasingly being integrated into educational settings to enhance personalized learning experiences. One study explored the use of emotion detection systems to identify students' emotional states in classrooms, allowing teachers to adjust their teaching strategies in real-time. The study utilized facial recognition and physiological signals to assess student engagement and emotional responses. The results showed that personalized learning interventions based on real-time emotion recognition led to improved student performance and engagement (Jones & Smith, 2023).

Another study used a method using adaptive windows and extracted fine-grained features based on eye movement. This research surrounded the MOOC learning environment. This approach enhances sample quality by segmenting data according to emotional fluctuation cycles and extracting temporal relationships. It achieved a 65.1% accuracy for four-category emotion recognition using eye movement signals.

2.10 Research Gaps

While significant progress has been made in the development of emotion recognition systems there remain several key research gaps as highlighted below

i Limited Data Representation for Neurodiverse Populations

A significant gap in the current literature and existing emotion recognition systems is the lack of diverse datasets representing neurodiverse populations, particularly children with Down syndrome. Most publicly available datasets are biased toward typically developing individuals. This gap impacts the generalization of emotion recognition models when applied to children with Down syndrome, whose facial expressions may deviate from typical patterns due to the distinct features associated with their condition.

ii Small and Imbalanced Datasets

The limited amount of labeled data not only affects the training process but also exacerbates challenges like class imbalance. While some studies have used data augmentation techniques or synthetic data generation, they remain inadequate in addressing the overall dataset size, which remains a significant challenge. Larger and more representative datasets are needed to build robust models capable of reliably predicting emotional states across all emotional classes.

2.11 Conceptual framework

The solution is conceptualized to work in the following manner. First, we would capture facial expressions from individuals during therapy sessions. For this study, pre-existing datasets of facial expressions were utilized, and no primary data collection is required. The data underwent preprocessing steps to enhance the quality of the images, including noise reduction and image enhancement. Once the data is cleaned and optimized, feature extraction techniques were applied to determine the most relevant features for emotion classification. The extracted features were then used to train and develop classification models designed to categorize emotions, specifically targeting positive,

negative, or neutral emotional reactions.

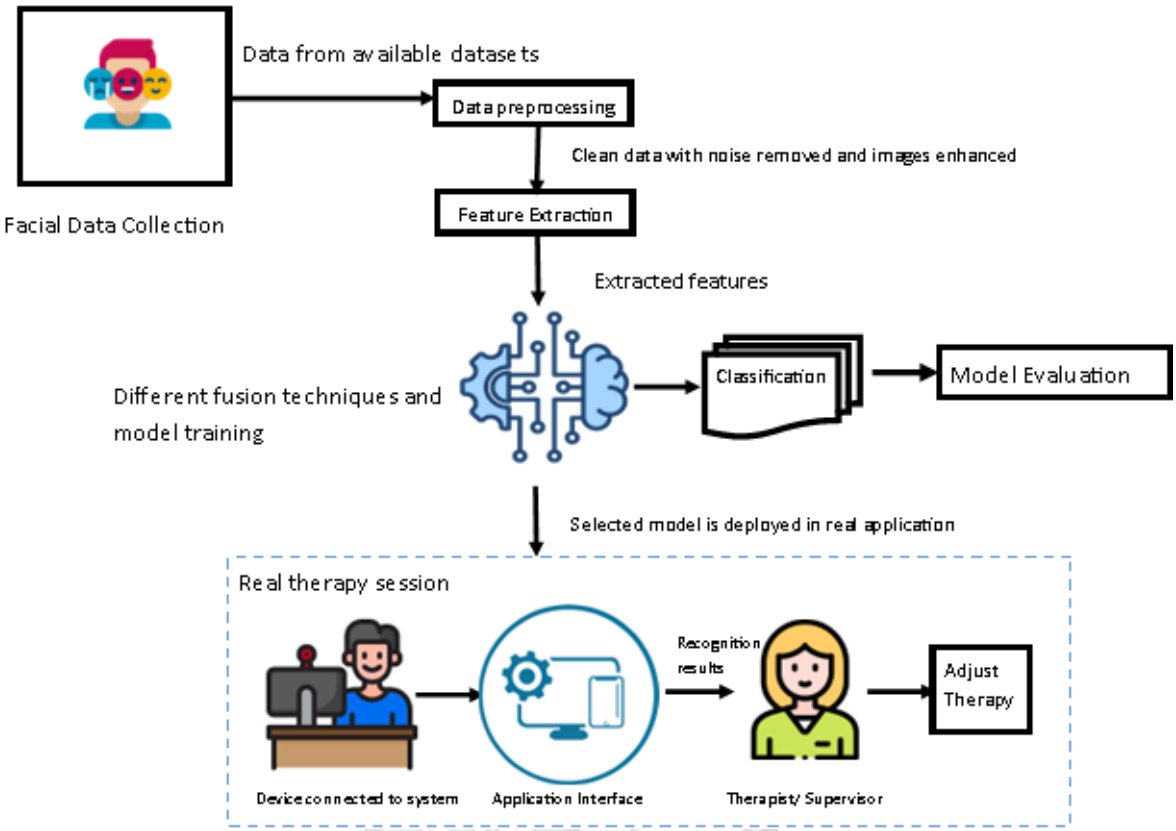


Figure 2.9 Conceptual Framework

Once the models are developed, users, such as therapists or caregivers, are required to utilize a webcam to capture facial expressions. During a therapy session, the system recorded the video feed, which is then preprocessed, sent to the emotion recognition model, and analyzed in real time. Based on the analysis, the system provided the recognition results, which the therapist could use to adjust the therapy, accordingly, ensuring that the emotional state of the individual was effectively monitored and responded to during the session.

Chapter 3 Research Methodology and Design

3.1 Research design

The research design refers to the overall strategy chosen to integrate the different components of the study in a logical way that addresses the research question. It is the essential blueprint for the collection, measurement, and analysis of relevant data. Being that there are several designs, the criterion of selection is based on depends on the nature of the question and the availability of resources.

This study employed an Applied Research Design, which is suitable given the practical goal of developing a functional emotion recognition system for children with Down Syndrome in therapeutic settings. The applied nature aligns with the project's objective: creating a real-world tool to assist in dolphin-assisted therapy by detecting emotional states through facial data. This design ensures the research moves beyond theory into practical utility, focusing on implementation and evaluation in a specific context.

To ensure practical effectiveness, this research explored different models and approaches to assess their outcomes in a real-world therapy setting. Additionally, an active learning approach was employed, where the most relevant and optimal data points were selected for training. Finally, multiple algorithms were rigorously evaluated to determine their effectiveness, leading to the selection of the most suitable model for deployment in therapy sessions. (Kamiri & Mariga, 2021a)

3.2 Methodological Approach

The research methodology answers the question "How." It outlines the systematic approach to research in a way that ensures valid and reliable results and addresses the objectives.

For this research, an Agile methodology was followed to maximize iterative and progressive development. The specific agile methodology adapted was the Rapid Application Development (RAD), which allowed small but functional increments of a project through sprints.

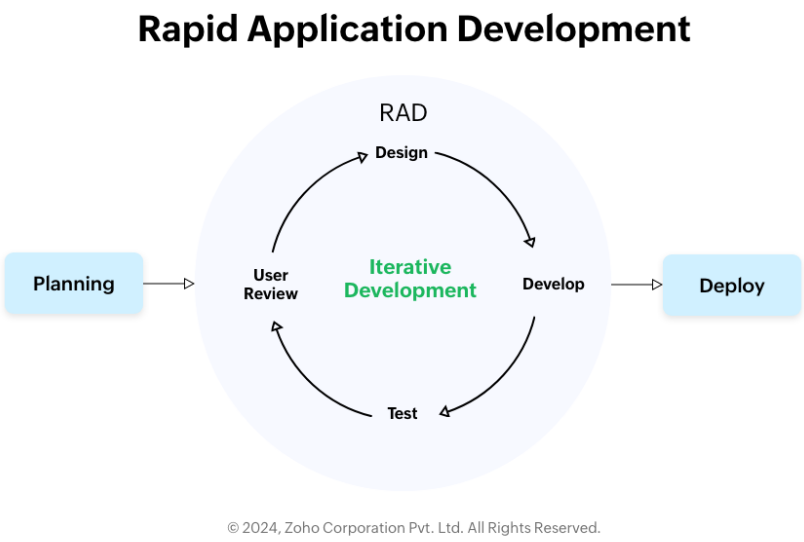


Figure 3.1 Rapid Application Development Design

This dynamic environment created room for adaptation to changing requirements and allowed for feedback incorporation throughout the process.

RAD makes extensive use of prototyping to get a basic, functioning version of the software into the hands of consumers as early as possible. In addition to this, RAD facilitates a more efficient use of resources since the focus was on creating a prototype and collecting feedback.

3.2.1 Planning

The first stage of the cycle began with defining the project requirements. This entailed the engagement of all stakeholders to determine the goals, the timelines and strategies to address any limitations that may arise. This stage ensured that everyone is on the same page in terms of expectations and deliverables. Within the Rapid Application

Development Methodology, there is an allowance of change in requirements at any point within the cycle and this was considered. For this research, this stage also included defining the objectives, ang gathering requirements such as identifying available DS dataset.

i. Algorithm Selection

This stage involved choosing appropriate algorithms, models and the consequent metrics for evaluating model performance. For a standardized comparison and proper understanding of which model best suited our facial emotion recognition task, we employed multiple algorithms: CNN, SVM, Random Forest (RF), DNN, and MLP; each was selected based on its strength in handling different aspects of our data as identified in the literature review. Since the dataset comprised facial images as well as structured features like facial landmarks and Action Units (AUs), it was essential to test models capable of learning from both spatial and numerical data types. Convolutional Neural Networks (CNNs) were ideal for the spatial data due to their strength in learning spatial hierarchies

In contrast, SVM and Random Forest were applied to the extracted facial landmark coordinates and AU values, which were more structured (numerical) and tabular in nature. SVM is well-suited for high-dimensional, small-to-medium datasets and excels in binary or multiclass classification tasks. RF, a powerful ensemble method, was chosen for its robustness and ability to handle non-linear relationships without heavy preprocessing. Deep Neural Networks (DNNs) and Multi-Layer Perceptron (MLPs) were included to test how well generic feed-forward architectures could model both image and numerical data when trained with sufficient depth and neurons. By comparing these, we aimed to determine which architecture best balances accuracy, generalizability, and interpretability for emotion recognition in children with Down Syndrome.

ii. Evaluation metrics

We used the standard evaluation metrics such as confusion matrix, accuracy, precision, and recall to comprehensively evaluate our .While accuracy provides a general measure of how often the model predicts correctly, it can be misleading when dealing with imbalanced .Precision and recall give a deeper insight into the model's performance on each class while recall shows how many actual positive cases were captured (minimizing false negatives. The confusion matrix allowed us to visualize the model's strengths and weaknesses per class and supported more informed decisions during tuning and model comparison.

We also selected the F1-score due to the class imbalance present in our dataset. Accuracy can be misleading in such scenarios. The F1-score, which is the harmonic means of precision and recall, offers a more reliable metric by ensuring both false positives and false negatives are penalized.

3.2.2 Prototyping

The second phase is prototyping. This critical phase where requirements are modelled and transformed into practical, testable implementations. In this research, prototyping was used to iteratively build and refine emotion recognition models available data objectives. Prototypes allowed for the inclusion of user feedback and the incremental adjustment until the project's goals are accomplished.

3.2.2.1 Baseline Model Development

As guided by the scope of the research outlined in the planning phase, the next step was building the initial model / prototype. The aim of a prototype was to produce a working design that can be demonstrated. For this research we began with the development of a baseline model which entailed training a custom CNN without pre-trained models for

benchmarking. As follows, we only utilized the dataset for Down Syndrome children and applied standard data preprocessing steps such as normalization, resizing and augmentation.

Most research in machine learning is mainly quantitative and follows standardized statistical approaches. Similarly, this research used a quantitative approach and adopted a general design for each model development as shown below. (Kamiri & Mariga, 2021b)

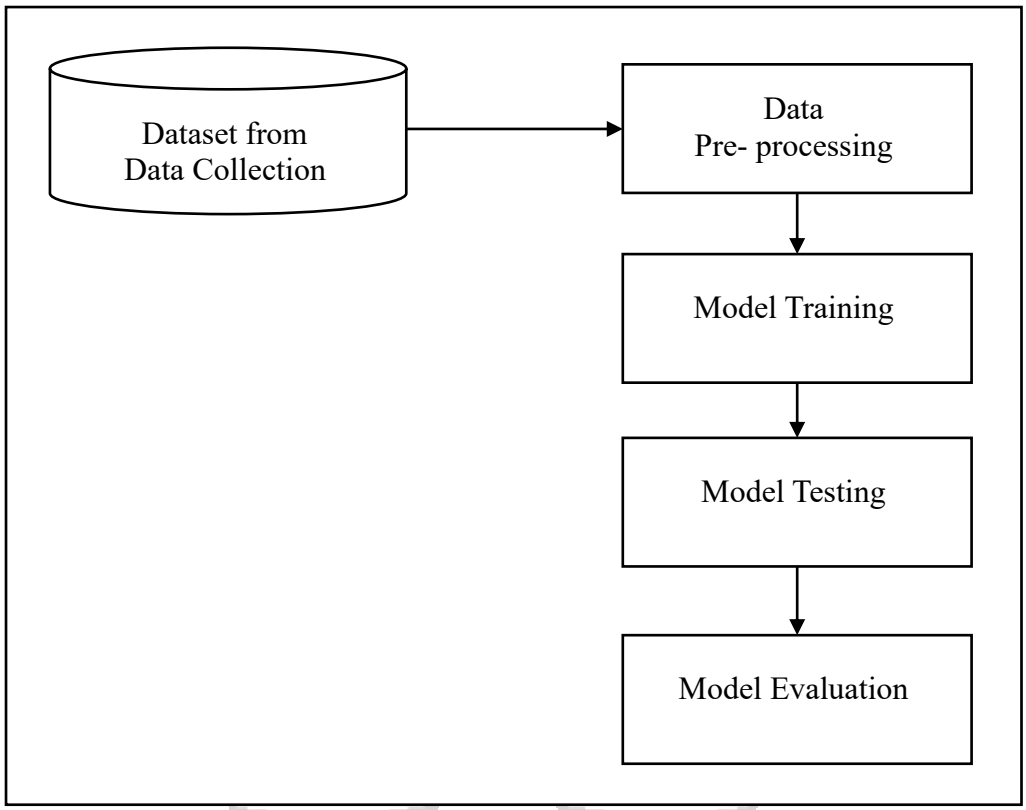


Figure 3.2 General Methodology for Machine Learning Projects

i Data Collection

This stage involved the collection of data either primary or secondary. This research utilized the custom DS Dataset. Specifically, the DS dataset includes facial images of subjects taken during experiments and are designed to elicit various emotions. This data was labelled by professional psychologists and was gathered in a controlled environment. (Moreno Escobar et al., 2023).

ii Data Pre-processing

This is a crucial stage that involves cleaning, normalization, noise reduction and feature selection. This stage is especially important for this research due to the different modalities involved. The feature selection and combination directly influence the performance of the models. We used PCA for selecting the optimal features as it identifies correlation among various attributes of a dataset.

iii Model Training and Testing

This stage was reliant on the choice of algorithms. All the models were trained in data obtained. The data was split into train and test data. Techniques such as cross validation were used for training. Finally, the models were evaluated on the test data.

iv Model Evaluation

The model was evaluated based on certain parameters such as true positives, true Negative, False Positive, False negative (from confusion matrix) and derived measures such as Accuracy, precision, recall and F1-Score.

3.3Experimentation and Iteration

The models were developed using different architectures, activation functions and optimizers. This also included fine-tuning of the hyperparameters (learning rate, batch size and dropout rate). Finally, each model was evaluated, and comparisons were drawn.

3.3.1 Deployment Phase

The deployment phase included integrating the working prototype to the real-world environment as well as evaluating generalizability on unseen test data. The working model was optimized for a real-world application.

3.4 Data

3.4.1Population/ Sampling

The target population of this research was children diagnosed with Down syndrome, specifically within the age range of 5 to 17 years. The sampling technique used was convenient sampling because secondary datasets were accessible to the researcher. This group was selected due to the increasing need for adaptive, non-invasive technologies that can support emotional communication and therapeutic monitoring for individuals with cognitive and developmental challenges. The system was designed to observe, classify, and summarize emotional states during therapeutic activities, with a focus on Dolphin-Assisted Therapy (DAT) as a real-world application context.

3.4.2 Data collection

Secondary data was utilized for this research due to the challenges around primary data collection. The study employed secondary data collected from a specialized Dolphin-Assisted Therapy program in Mexico. The publicly available dataset includes labelled facial images of three children with Down syndrome during their therapy sessions. These images capture a range of emotional expressions relevant to the context of therapeutic engagement. The dataset is already organized into three subsets for the purpose of machine learning experimentation:(Moreno Escobar et al., 2023).

i Training Set: Used to train the CNN and hybrid classification models.

ii Validation Set: Used to fine-tune hyperparameters and avoid overfitting during model training.

iii Test Set: Used for final evaluation of model performance and generalizability.

Each subset contains images labeled according to emotional states such as attention, dislike, calm, and surprise, enabling supervised learning and evaluation of classification accuracy.

3.5 Utilization of Research Results

The main purpose of this study is to enhance emotional insight and therapy personalization for children with cognitive impairments, particularly those with Down syndrome undergoing Dolphin-Assisted Therapy (DAT). The proposed emotion assessment system aims to complement therapeutic efforts by enabling data-driven, adaptive support in real time. The following outlines how the results of this study were utilized:

i **Integration into Therapy Environments:**

The primary output of this research is a multi-modal emotion recognition system using facial features extracted through CNNs, Mediapipe landmarks, and DeepFace emotion scores. This system is intended to be deployed during real-time therapy sessions to assist caregivers in interpreting the child's emotional state. The model's outputs can inform adaptive adjustments during therapy, such as modifying stimuli when negative emotions are detected.

ii **Clinical and Research Applications**:

The findings may contribute to the growing body of clinical research focusing on behavioral cues in children with developmental disorders. By offering an automated, non-invasive emotion detection tool, this system can support clinicians in assessing therapy effectiveness over time, tracking emotional engagement, and identifying stress patterns.

3.6 Dissemination of research results

i **Academic Publications**

The research findings will be submitted to peer-reviewed journals in fields such as affective computing, special needs education, therapeutic technology, and computer vision. Publication will contribute to scholarly discourse on the intersection of AI, healthcare, and developmental therapy.

ii **Stakeholder Collaboration**

Efforts were made to collaborate with special education institutions, child psychologists, and non-profit therapy centers to pilot the tool. Their feedback is vital for refining the tool and ensuring real-world applicability.

3.7 Ethical Considerations

The research only used data that has been collected and processed in line with the relevant ethical guidelines, especially given the sensitivity of working with a vulnerable population. With regards to data privacy, personal identifiable information was withheld or obscured to the public throughout the study.

Chapter 4 System Analysis and Design

4.1 Introduction

System Analysis and Design (SAD) is a crucial phase in software development which focuses on understanding user requirements and modelling or designing the solution to ensure the implemented system meets the research objectives. In the context of this study, this chapter provides the foundation for identifying the technical and functional expectations of the emotion recognition system and how it could be structured to support therapeutic decision-making in children with Down syndrome. This chapter therefore presents both the functional and non-functional requirements derived from the system goals and intended use.

For the design approach, this research adopted the principles of Object-Oriented Analysis and Design (OOAD). OOAD was particularly well suited for this system as it promotes modularity, reusability, and abstraction—core attributes necessary in building scalable and extensible software components Through OOAD, the system was broken down into well-defined objects that map directly to real-world entities and actions within the therapy setting.

The blueprint of the system is presented through a series of design diagrams, including case diagrams, class diagrams, sequence diagrams, and architecture diagrams. These visual tools capture how users interact with the system, how internal components communicate, and how data moves across different modules. In addition to architectural modelling, the chapter concludes with a review of the key technologies and platforms used in developing the prototype, including machine learning frameworks, facial analysis libraries, and web-based deployment tools.

4.2 System Requirements

System requirements are descriptions of what the system should do and its operational constraints (how it should perform). These requirements are categorized as functional and non-functional characteristics.

4.2.1 Functional

The functional requirements outline the specific operations the system must perform to achieve its goal of emotion recognition during therapy. These requirements reflect the logical flow of data from input to feedback.

i The system should capture facial input from either uploaded video files or live webcam streams.

ii The system should preprocess incoming frames to ensure quality and consistency.

iii The system should extract relevant features from each frame.

iv The system should classify the input into predefined emotional categories.

v The system should provide emotional recognition results with minimal delay.

4.2.2 Non-Functional

The non-functional requirements define the quality attributes and constraints that the system must adhere to. These ensure that, beyond simply functioning, the system performs efficiently, remains usable, and meets expectations.

i The system should provide near-time processing and feedback.

ii The system should maintain high reliability and robustness across diverse environmental conditions.

iii The system should be scalable and modular.

iv The system should offer high usability and accessibility for therapists and non-technical users.

v The system should be compatible with standard hardware.

4.3 System Narrative

The emotion recognition system was designed to support Dolphin Assisted Therapy by automatically analyzing facial expressions captured during therapy sessions involving children with Down syndrome. Users interact with the system by uploading a recorded therapy session or activating a live webcam stream during therapy. The system captures facial data from the child and processes it using a combination of deep learning and facial analysis techniques.

These predictions are updated frame by frame and aggregated to determine dominant emotional trends during the session. A rule-based feedback engine interprets these trends and generates guidance, allowing therapists to adjust their approach dynamically.

For live sessions, the system processes every nth frame (e.g., every 10th frame) to balance real-time performance with accuracy. A rule-based feedback module monitors the emotional trend over time, identifying prolonged negative or positive responses and triggering contextual feedback.

4.4 System Analysis Diagrams

This section contains the blueprints of the different modules within the system and how they interact architecture and workflows to provide a clear understanding of its structure and interactions.

4.4.1 Sequence Diagram

A sequence diagram shows the interaction between a user and the core components of a given system. The process begins when the user initiates the system through the user interface, either by uploading a video or activating a live webcam stream. The system then

records the input and passes it to the Video Handler, which is responsible for extracting individual frames from the session. These are what we are interested in capturing.

Each extracted frame is then sent to the next component which is the Landmark Extractor, where facial landmarks are detected using the Mediapipe framework. These landmark features are embedded and sent along with the original frame to the Deep Face API, which extracts Action Unit (AU) scores representing emotional intensity and facial expressions. The output includes both the spatial landmark features and the AU-based emotion features.

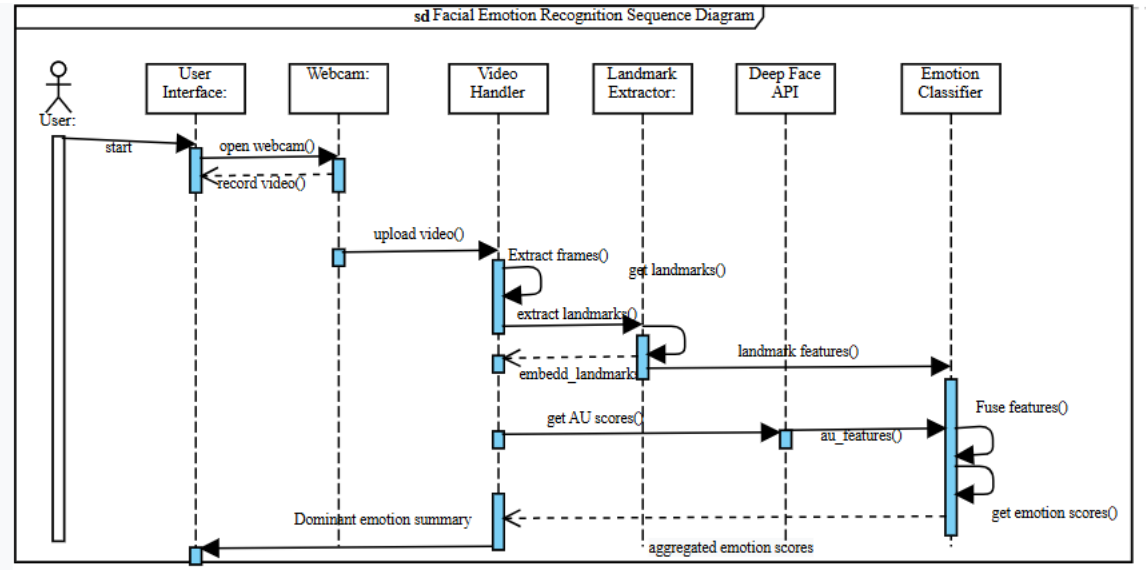


Figure 4.1 System Sequence Diagram

These features are then passed to the Emotion Classifier, which fuses the multimodal inputs and performs emotion prediction on each frame. The resulting emotion scores are aggregated across the video session to determine the dominant emotional state. This information is then returned to the user via the interface as a summarized emotional report, supporting the interpretation of emotional trends throughout the therapy session.

This sequence highlights the modular design and step-by-step orchestration of components, ensuring efficient, explainable, and real-time emotion recognition for therapeutic support.

4.4.2 Use Case Diagram

The use case diagram illustrates the primary interactions between users and the facial emotion recognition system. There are two main actors: the Subject (the child undergoing therapy) and the Therapist, as well as a supporting system actor labeled as Model, which represents the trained machine learning model.

The process begins when the Subject starts the webcam, triggering the system to record a video of the therapy session. This process is supported by the Webcam hardware. The recorded video is then used to extract frames, which serve as the basis for analysis. These frames pass through the Feature Extraction process, which includes the use of Mediapipe to detect facial landmarks and the DeepFace API to compute emotion-related Action Units. All extracted features are then combined in the Feature Fusion step.

Once features are fused, the Model is invoked to perform Emotion Classification, determining the emotional state of the subject in each frame. The results of this classification are then displayed to the Therapist through the user interface. The Display Result use case can optionally extend to Adjust Therapy, allowing the therapist to respond dynamically to the subject's emotional state based on real-time or summarized feedback.

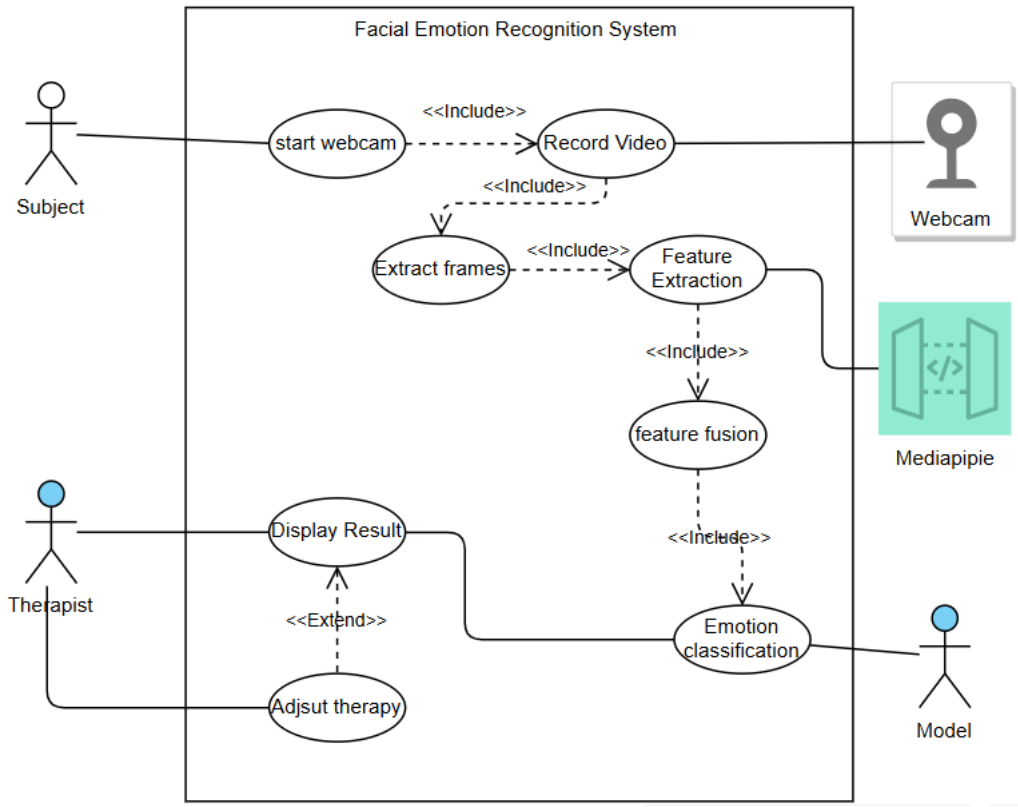


Figure 4.2 Use Case Diagram

4.4.3 Flow chart

The interaction flow begins when the user (therapist or caregiver) uploads a video through the front end or starts a recording. This video is sent to the backend via HTTP requests, where it undergoes emotion recognition processing. Facial landmarks among other features are extracted and computes emotion scores from the video. The results are then stored in the database, which can be accessed later for further analysis or reporting. Once the data is processed, the analysis results are presented. Based on the analysis, the therapist may trigger real-time adjustments to the therapy process.

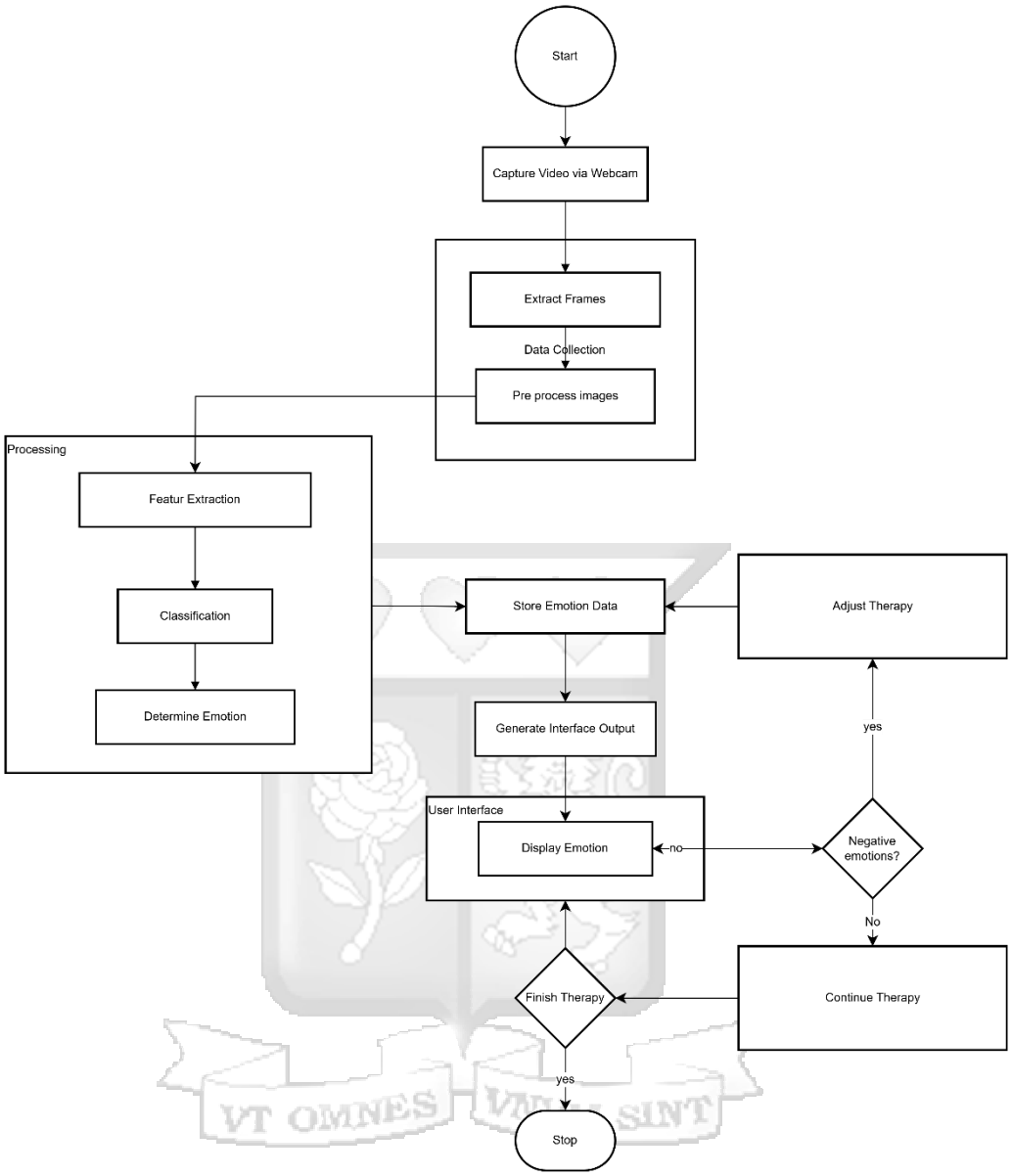


Figure 4.3 Flowchart of System

4.4.4 Entity Relationship Diagram

The entity relationship diagram (ERD) provides a structural overview of the database design. The tbl_subject table stores the attributes of each subject, such as age and any relevant notes which is not confidential. This table has a one/many to many relationship with the tbl_video which stores the recorded sessions. The attributes we are interested in are the ,logs and metadata such as the upload time and file name. The relationship between subjects and videos is one-to-many, as a single subject can undergo multiple therapy

sessions over time. The individual frames that are extracted from the video feed are stored in the tbl_frame table. Each frame is also timestamped and linked back to the corresponding video with a secondary key. This frame-level breakdown is important because its enables a complete analysis of emotional state progression throughout the session.

The features extracted by the Landmark extractor are stored in the table tbl_feature The attributes are the facial landmarks and Action Unit (AU) scores obtained via Mediapipe and DeepFace, respectively. These features are the expected inputs to the trained classifier (the CNN model). which outputs predictions. The predictions for each frame are stored in the table tbl_prediction. The confidence score s also stored in this table so that both quantitative evaluation and summarization is made possible. Finally, the tbl_feedback table captures feedback from the system and the therapist as well. This table stores both timestamped emotional state and text-based feedback that is related to a unique therapy session for each user/subject and is intended to support the therapist's decision-making process during or after a session.

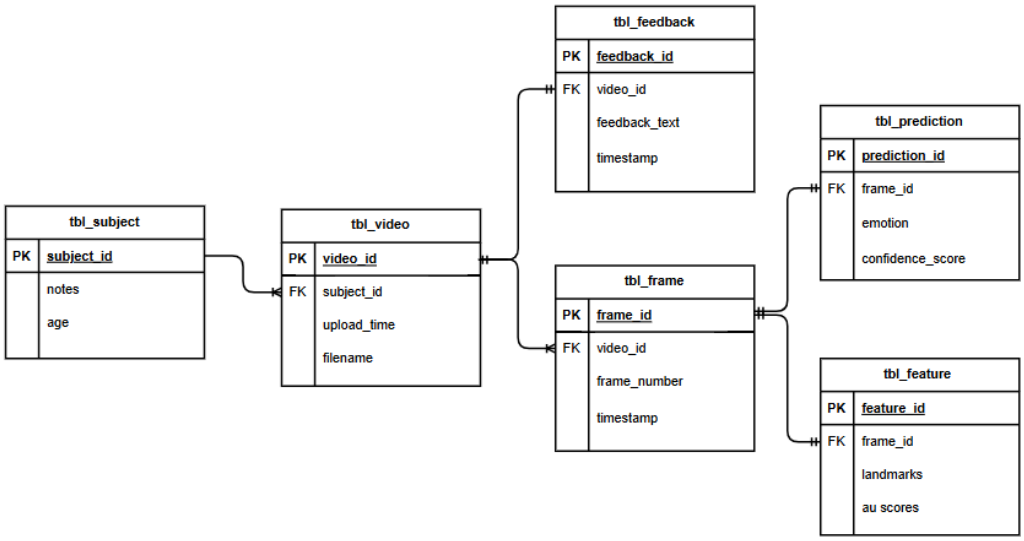


Figure 4.4 Entity Relationship Diagram

4.5 System Design Diagrams

4.5.1 Model Architecture

The baseline model employed in this study was a lightweight Convolutional Neural Network (CNN) specifically designed to classify facial images into four emotion categories. The process begins with an input image of shape 150 × 150 × 3, which is first passed through a rescaling layer to normalize pixel values to a [0, 1] range. This normalization is essential to stabilize the training process and accelerate convergence.

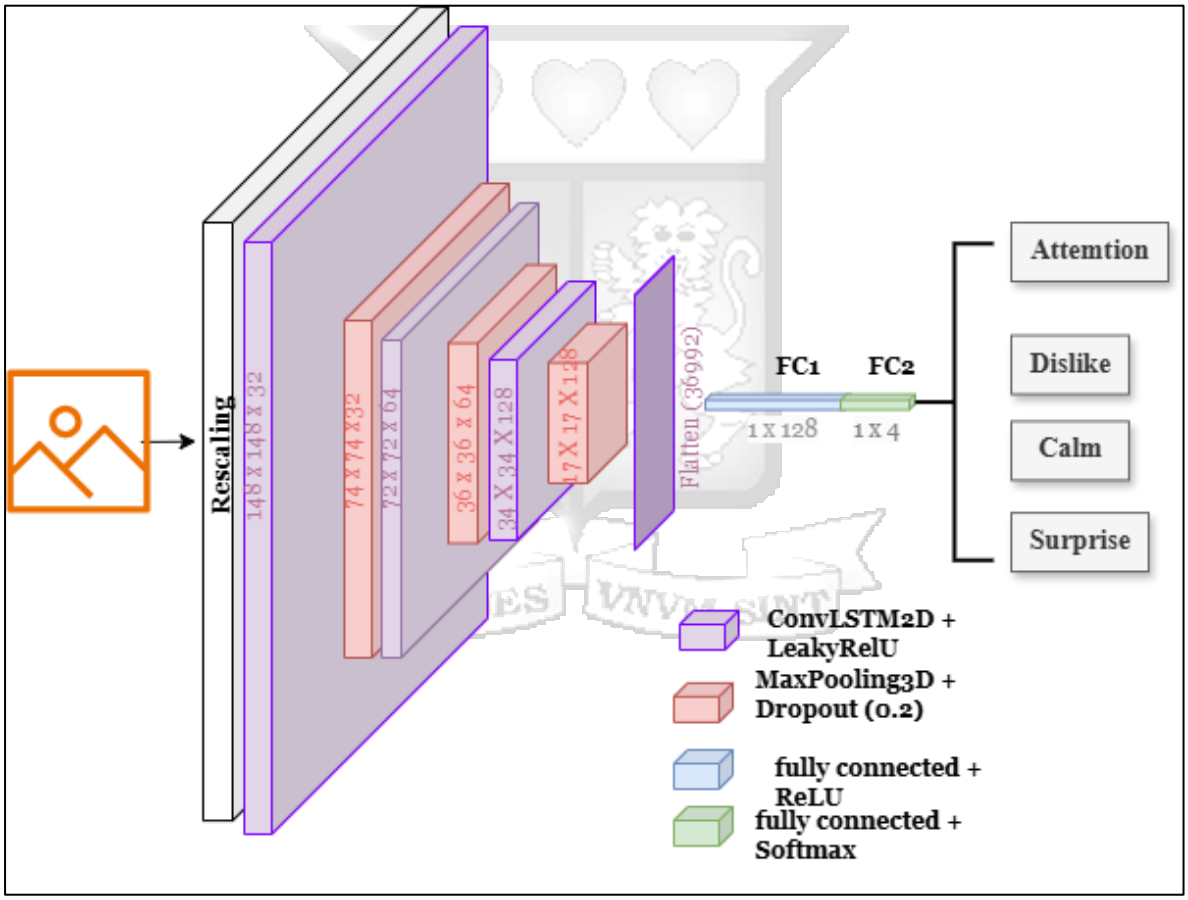


Figure 4.5 Custom CNN Architecture

The feature extraction component consists of three convolutional layers followed by max-pooling layers. Each convolutional layer progressively increases the number of filters while reducing the spatial dimensions of the feature maps:

i The first convolutional block uses 32 filters of size 3×3, producing an output of shape 74 × 74 × 32.

ii The second block doubles the depth to 64 filters, yielding a 36 × 36 × 64 output.

iii The third block increases depth to 128 filters, with the output feature map sized 34 × 34 × 128.

iv A final max-pooling operation reduces this to 7 × 17 × 128.

Following convolutional processing, the 2D feature maps are flattened into a 1D allowing the model to transition from feature extraction to classification. This flattened vector is passed into a dense (fully connected) layer with 128 units and ReLU activation, which enables non-linear transformations. A dropout layer is applied here (rate = 0.3) to mitigate overfitting by randomly deactivating neurons during training.

The final classification is performed using a dense output layer with 4 units, corresponding to the target emotion classes. A SoftMax activation function is applied to generate a probabilistic distribution over the classes, ensuring that the model produces interpretable predictions.

4.5.2 System Architecture

The whole system follows a simple web Application architecture that is depicted below which constitutes the frontend and the backend. The user interacts with the UI and the Requests are made to the web server which fetches data from the database and returns it to the user in line with the defined business and application logic..

The frontend is the user-facing part of the system, designed for interaction between therapists and the emotion recognition application. It is developed using standard web technologies, such as HTML, CSS, and JavaScript.

The backend consists of the core logic that handles the processing of input data (in this case, video files), emotion recognition, and session management. The backend is powered by a web server framework and handles communication with both the frontend and the database. Upon receiving an uploaded video from the user, the backend processes the video to detect facial features and extract emotion-related information. This processing step is reliant on the trained convolutional neural networks (CNNs) for the image recognition part and DeepFace for identification of the action unit.

To allow for a more responsive and adaptive experience, the backend is designed to manage any feedback or adjustments that might be required based on the emotion detected. For example, if a detected emotion falls outside of the desired range (such as having a neutral effect , instead of calming effect), the system can trigger something to alert the therapist such that the therapy process is adjusted in near real time or real time.

Lastly, The database serves as the storage layer, storing key information related to video sessions, emotion detection results, and feedback logs. This system is reliant on the relational database MySQL as the data is structured.

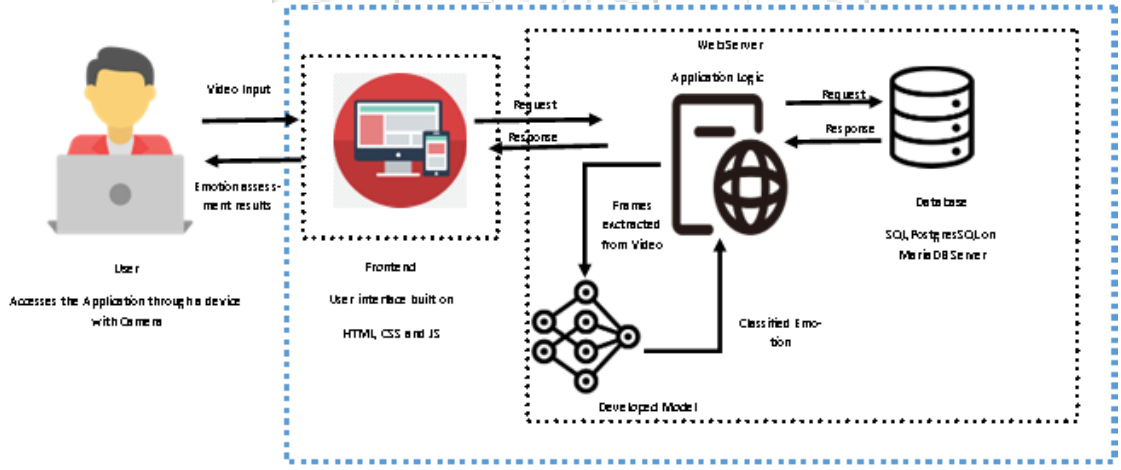


Figure 4.6 System Architecture

4.6 Tools and Techniques

4.6.1 Coding environment

All model development and experimentation were implemented using Google Colaboratory (Colab) a cloud-based Python environment that enables GPU acceleration for faster training. Colab was chosen due to its support for interactive notebook development and ease of integration with external files such as Google Drive for model checkpointing and dataset access.

The programming language used was Python 3.11, with the following key libraries integrated into the development pipeline:

i TensorFlow/Keras: Used for building and training the CNN and MLP models.

ii NumPy and Pandas: For handling numerical operations and data preprocessing.

iii Matplotlib and Seaborn: For visualizing training results, metrics, and model performance.

iv Scikit-learn: Used for model evaluation (confusion matrix, classification report), scaling, PCA, and classical models like SVM and LightGBM.

v OpenCV: Employed for video handling, frame extraction, and image preprocessing.

These tools were essential for maintaining a reproducible workflow and conducting rigorous model experimentation.

4.6.2 Feature extraction

Two primary sources of feature extraction were used to capture diverse representations of emotion:

i MediaPipe: A lightweight, real-time face mesh solution by Google, used to extract 468 facial landmarks per frame. These landmarks represent structural deformations and subtle muscle movements, providing essential spatial features.

ii DeepFace: This high-level Python framework integrates pre-trained facial recognition and analysis models. In this study, DeepFace was employed to extract Action Units (AUs) and emotion scores following the Facial Action Coding System (FACS) framework, allowing the system to infer underlying affective states like surprise, anger, and sadness.

Chapter 5 System Implementation and Testing

5.1 Introduction

This chapter details the development of various modules that were implemented for this study.

5.2 Description of Implementation Environment

It provides an outline of the hardware and software identifications that are required for the system to be implemented and to be fully operational.

5.2.1 Hardware Specifications

The most significant factors that were considered while determining the hardware requirements have been summarized below:

Table 5.1 Hardware Specifications

Item	Minimum Specification	Recommended Specification
Processor	2 x 1.6 Ghz	4 x 1.6 Ghz
RAM	4 GB	6 GB
Hard Disk Storage	60+ GB of free space or more is recommended (Dependent on whether video files are saved locally or on the cloud)	
Input devices	Standard keyboard, mouse and Web Camera	
Output devices	High resolution monitor	

5.2.2 Software Specifications

i. Operating System

The system was designed to run in a standard Linux operating system. The system was developed within the Ubuntu 21.04 Linux based operating system, x64-based processor. Therefore, it is possible to run the system in a Linux environment.

ii. Web Server

The system was designed to work with Apache Web Server version 2.x. There are no special Apache configurations required to run the System.

iii. Database Server

The system was designed to work with a relational database management system. It was designed to be compatible with MySQL 5.7+ and PostgreSQL 9.6+. For the system to run efficiently certain grant privileges are required to be active. These are the following: INSERT, SELECT, UPDATE and DELETE.

5.3 Model Implementation and Testing

This section presents the three models developed in this study: The baseline convolutional neural network (CNN) model trained solely on raw facial images, the second introduced a feature-based approach, utilizing semantic emotion scores from DeepFace and geometric facial landmarks from Mediapipe. Lastly, multimodal fusion model combining raw image features extracted from a pre-trained CNN, semantic features (emotion scores), and geometric features (facial landmarks) were trained and evaluated.

All the models were evaluated using precision, recall, F1-score, and confusion matrices to assess their effectiveness in classifying four emotional categories:.

5.3.1 Dataset description

The dataset used in this study comprises 1,445 labeled facial images of children with Down syndrome, categorized into four emotional states: Surprise, Calm, Dislike, and Attention. These images were captured during Dolphin-Assisted Therapy (DAT) sessions and reflect spontaneous facial expressions observed in therapeutic settings. The dataset was divided into three predefined subsets: 933 images for training, 128 for testing, and 384 for validation. (Moreno Escobar et al., 2023).As shown in Table 5.2, the distribution across classes varies, with *Surprise* being the most represented (449 images) and *Attention* being the least (269 images). Each emotion class was stored in a dedicated directory, and the images are in .jpg format.

Table 5.2 DS Dataset Image Distribution

	Training	**Testing**	**Validation**	***Total***
Surprise	321	32	96	**449**
Calm	189	32	96	**317**
Dislike	289	32	96	**417**
Attention	134	32	96	**269**
Total	**933**	**128**	**384**	

5.3.2 Data pre-processing

For the preprocessing phase, images were first manually extracted from the compressed source and organized into separate folders corresponding to their respective emotion classes. As all images were in .jpg format and of manageable size, minimal structural cleaning was necessary. However, to enhance model generalization, increase data variability and address overfitting—particularly due to the limited dataset size—a set of

image augmentation techniques was applied before and during training. We used oversampling to balance the under-represented classes.

```
# Image Data Generator (augmentation settings)
augmentor = ImageDataGenerator(
    rotation_range=15,
    zoom_range=0.1,
    horizontal_flip=True,
    fill_mode='nearest'
)
```

Figure 5.1 Data Augmentation before training.

These transformations included horizontal flipping, rotation, zooming, and brightness adjustments, implemented through TensorFlow's ImageDataGenerator and Keras preprocessing utilities and were only applied to the training set.

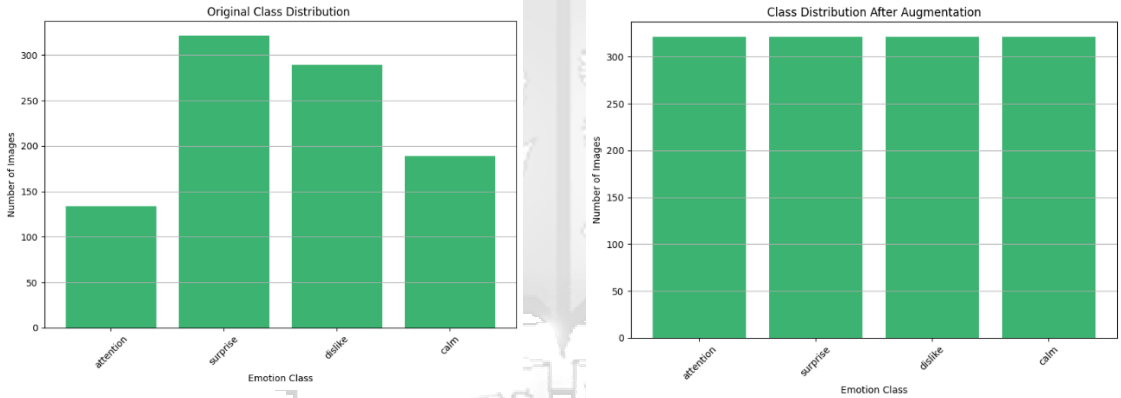


Figure 5.2 Comparison of Dataset before and after data augmentation.

There was no need to rescale to a fixed dimension as all the images were already 150 by 150. During training, the pixel values were normalized to the [0, 1] range to ensure consistency during model input.

5.3.3 Feature engineering

Feature engineering involved extracting and creating distinct features from raw image data, enhancing the model's ability to identify and classify emotions accurately.

The first step in feature engineering involved extracting key facial features that could be used to determine emotional states. This process was performed using two primary methods:

Facial Landmarks Extraction with MediaPipe:

MediaPipe, a machine learning library for computer vision, was used to detect and extract facial landmarks from the input images. These landmarks represent various key points on the face, such as the eyes, nose, and mouth, which provide essential spatial information about facial expressions. The raw image data was processed through MediaPipe to generate landmark coordinates, which were flattened into a single vector per image, creating a feature set for further analysis.

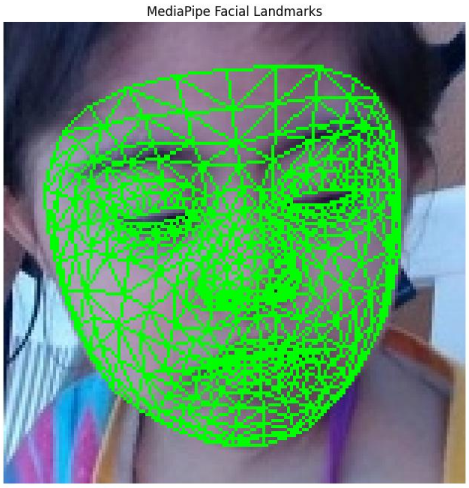


Figure 5.3 Sample Landmark Mapping by Mediapipe

Action Unit (AU) Extraction with DeepFace:

DeepFace, a deep learning-based face analysis library, was employed to compute Action Unit scores, which represent different facial muscle movements associated with specific emotions (e.g., the movement of the eyebrows for surprise). These AUs were derived from each image, providing additional emotion-relevant features. These

features were returned as a vector, indicating the intensity of different emotional expressions.

Figure 5.4 DeepFace AU scores on sample image.

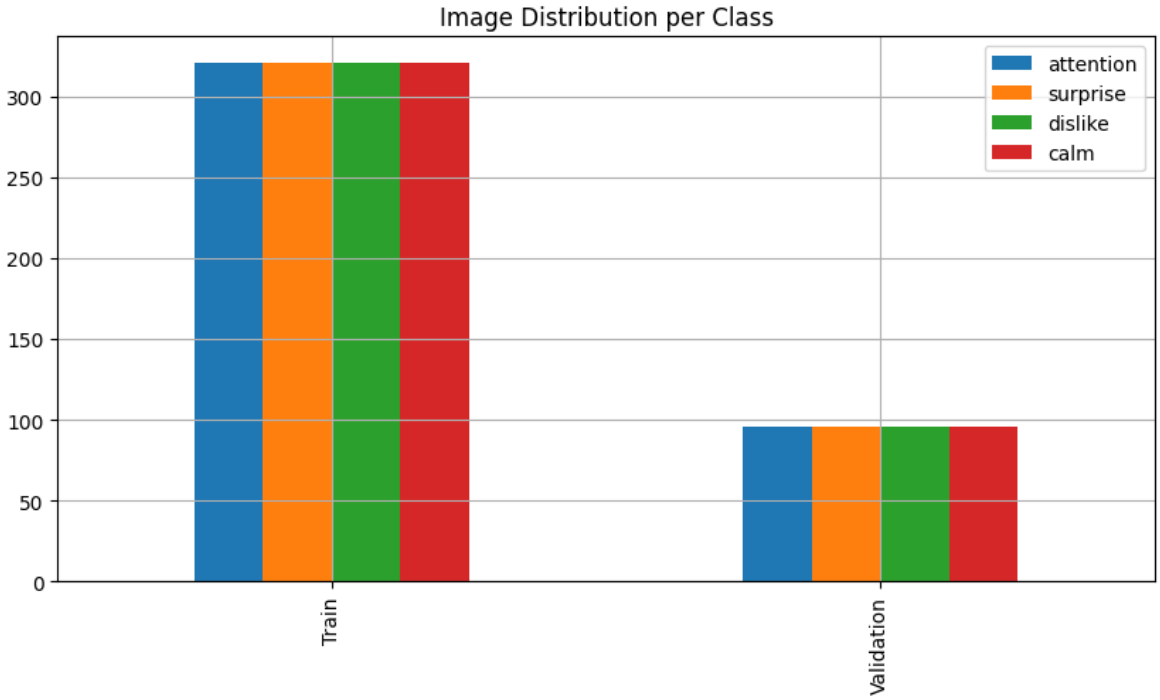


Figure 5.5 Class Distribution for Dataset used for Feature based model.

5.3.4 Image Based Model Development

For the first prototype, which was based on raw images, a simple custom Convolutional Neural Network (CNN) model was developed and trained using the balanced dataset as described in the preprocessing stage.

5.3.4.1 Model Training

Model tuning was carefully managed to prevent overfitting, given the relatively small dataset. Training was conducted in the TensorFlow/Keras environment using a custom CNN architecture with three convolutional layers followed by fully connected layers. To enhance generalization, in-network data augmentation techniques such as random rotation, horizontal flipping, and zooming were applied, and dropout layers were introduced to reduce overfitting. We used Grid Search for tuning parameters such as learning rate ([0.001, 0.0005]), batch size ([32, 64]), and number of convolution layers ([2, 3, 4]). The model was trained for 30 epochs with a batch size of 32 using the Adam optimizer and categorical cross-entropy loss. Early stopping was used to halt training when performance plateaued, while model checkpointing ensured only the best-performing model based on validation accuracy was saved.

```
data_augmentation = tf.keras.Sequential([
    layers.RandomFlip('horizontal'),
    layers.RandomRotation(0.1),
    layers.RandomZoom(0.1)
])

def build_cnn_model():
    model = models.Sequential([
        data_augmentation,
        layers.Rescaling(1./255, input_shape=(IMG_SIZE[0], IMG_SIZE[1], 3)),
        layers.Conv2D(32, (3, 3), activation='relu'), layers.MaxPooling2D(),
        layers.Conv2D(64, (3, 3), activation='relu'), layers.MaxPooling2D(),
        layers.Conv2D(128, (3, 3), activation='relu'), layers.MaxPooling2D(),
        layers.Flatten(),
        layers.Dense(128, activation='relu'),
        layers.Dropout(0.3),
        layers.Dense(len(class_names), activation='softmax')
    ])
    model.compile(optimizer='adam', loss='categorical_crossentropy', metrics=['accuracy'])
    return model

cnn_model = build_cnn_model()
```

Figure 5.6 Custom code for the CNN baseline model.

Throughout the training process, both accuracy and loss metrics were monitored and visualized to assess the model's learning progression. The final model was saved in .h5 format and served as the foundation for evaluating performance on the validation and test sets.

5.3.4.2 Model Evaluation

Following training, the best model (as determined by validation accuracy) was evaluated on the held-out validation set. The classification performance metrics are summarized as follows:

	precision	recall	f1-score	support
attention	0.51	0.81	0.62	96
calm	0.81	0.57	0.67	96
dislike	1.00	1.00	1.00	96
surprise	0.92	0.64	0.75	96
accuracy			0.76	384
macro avg	0.81	0.76	0.76	384
weighted avg	0.81	0.76	0.76	384

Figure 5.7 CNN Performance Per Class

Overall, the model achieved an accuracy of 0.76 % on the validation set, with a consistent macro and weighted F1-score of 0.76.

5.3.5 Feature Based Model Development

The second prototype focused on a feature-based approach for emotion recognition, where instead of relying solely on raw image data, we extracted facial landmarks (geometric features) and Action Unit (AU) emotion scores (semantic features) as structured features.

5.3.5.1 Model Training

On the feature set we trained a couple of shallow learning and deep learning algorithms. Random Forest (RF) and Support Vector Machine (SVM) were utilized due to their effectiveness in handling tabular data and their suitability for smaller datasets. Additionally, deep learning models like Multi-Layer Perceptron (MLP) and Dense Neural

Networks (DNN) were also explored for their ability to capture non-linear relationships and complex patterns within the data.

A Multi-Layer Perceptron (MLP) model was used to classify emotions based on the fused features. The model included two hidden layers with ReLU activation, batch normalization, and dropout regularization to mitigate overfitting. We did not need to address class imbalance, as this was addressed by oversampling techniques on the raw dataset. In some configurations, Principal Component Analysis (PCA) was optionally applied to reduce dimensionality and improve training efficiency. The training process was monitored using validation metrics and early stopping was employed to retain the best-performing model.

For MLP, training was performed over 30 epochs, using a batch size of 32 and the Adam optimizer, with categorical cross-entropy as the loss function. Early stopping was implemented to prevent unnecessary training once the model's performance stagnated/plateaued, and a model checkpoint callback was used to save the best-performing model at each epoch based on validation accuracy. Throughout the training process, both accuracy and loss metrics were monitored and visualized to assess the model's learning progression.

5.3.5.2 Model Evaluation

The classification performance metrics are summarized as follows for each model:

Random Forest outperformed SVM and the other deep learning algorithms when trained on selected features, particularly in the fused feature set.

MLP achieved an accuracy of 0.5 with a low F1 score of 0.38 and struggled to classify attention and surprise as shown below.

	precision	recall	f1-score	support
attention	0.00	0.00	0.00	32
calm	0.33	1.00	0.50	32
dislike	1.00	1.00	1.00	32
surprise	0.00	0.00	0.00	32
accuracy			0.50	128
macro avg	0.33	0.50	0.38	128
eighted avg	0.33	0.50	0.38	128

Figure 5.8 MLP Performance Per Class

SVM did better than MLP but also struggled with the Attention Class. It achieved an accuracy of 0.58 as shown below.

	precision	recall	f1-score	support
attention	0.41	0.12	0.19	96
calm	0.38	0.95	0.55	96
dislike	1.00	1.00	1.00	96
surprise	1.00	0.23	0.37	96
accuracy			0.58	384
macro avg	0.70	0.58	0.53	384
weighted avg	0.70	0.58	0.53	384

Figure 5.9 SVM Performance Per Class

5.3.6 Final Model Development

The last prototype explored a hybrid approach. These learned embeddings from the CNN model were combined with the features extracted from facial landmarks and AU scores. The fusion of these embeddings aimed to create a richer feature representation that would improve the model's ability to classify emotions.

5.3.6.1 Model Training

We used the Convolutional Neural Network (CNN) designed in the first prototype to extract features from raw image data. This was to get the embedding from the last layer before the output layer. CNN learns spatial hierarchies in the images and outputs a set of embeddings (or features) representing the image content. We then integrated the extracted

and preprocessed features from the second prototype we developed. So, the final model can be described as a hybrid model.

5.3.6.2 Model Evaluation

The model achieved the highest accuracy and F1-score of 0.84 as shown in Figure 5.10 below. The model achieved a precision of 0.82 for attention, 0.81 for calm, 1.00 for dislike, and 0.75 for surprise. Recall values were 0.72, 0.88, 1.00, and 0.78 respectively, indicating how well the model identifies true positives for each class as shown in Figure 5:10

	precision	recall	f1-score	support
attention	0.82	0.72	0.77	96
calm	0.81	0.88	0.84	96
dislike	1.00	1.00	1.00	96
surprise	0.75	0.78	0.77	96
accuracy			0.84	384
macro avg	0.84	0.84	0.84	384
weighted avg	0.84	0.84	0.84	384

Figure 5.10 Hybrid Model Performance Per Class

The F1-scores, which balance precision and recall, were highest for dislike at 1.00, followed by calm (0.84), attention (0.77), and surprise (0.77). Each category had equal support with 96 samples. Overall, the model achieved an accuracy of 0.84, with both macro and weighted average precision, recall, and F1-score also at 0.84, reflecting balanced performance across the classes on a total of 384 samples.

5.4 System Implementation and Testing

This section outlines the implementation process of the therapy system, which includes the development of both the backend and frontend components. It covers the integration of the emotion detection model into a live system, incorporating real-time emotion recognition, session tracking, and user management functionalities.

5.4.1 Backend Development

The backend was developed using Flask, a lightweight web framework for Python. The core functionality of the backend involves managing user sessions and interacting with the deep learning model to provide real-time emotion predictions. Flask-Login was used for user authentication, where therapists could log in, register, and manage their profiles. SQLAlchemy was utilized to manage the database and store user credentials, session data, and emotion classification. An example of the User and session model is seen in the Figure below.

```
38  # User model for authentication
39  class User(UserMixin, db.Model):
40      id = db.Column(db.Integer, primary_key=True)
41      username = db.Column(db.String(150), unique=True, nullable=False)
42      password = db.Column(db.String(150), nullable=False)
43      email = db.Column(db.String(150), unique=True, nullable=False)
44
45      def set_password(self, password):
46          self.password = generate_password_hash(password)
47
48      def check_password(self, password):
49          return check_password_hash(self.password, password)
50
51      def __repr__(self):
52          return f'<User {self.username}>'
53
54  # Session model for tracking therapy sessions
55  class Session(db.Model):
56      id = db.Column(db.Integer, primary_key=True)
57      therapist_id = db.Column(db.Integer, db.ForeignKey('user.id'), nullable=False)
58      session_start = db.Column(db.String(100))  # Start time of the session
59      emotion_detected = db.Column(db.String(100))
60      predictions = db.relationship('EmotionPrediction', backref='session', lazy=True)
61
```

Figure 5.11: User Model Code snippet from Backend.

Each therapy session was tracked in a session table, which included the therapist's ID, session start time, and the detected emotion. Each emotion prediction was linked to the corresponding session, ensuring that every frame processed by the model was tied to the correct session.

5.4.2 Real-Time Emotion Detection

For the emotion detection, we selected the model that performed the best which was the CNN model in 5.3.6 built using Keras. The model was loaded into the Flask application and used to predict emotions in real-time from video frames captured by the webcam. Figure 5.12 below shows a snippet of code with the function to set up the web camera.

```
7
8   // 3. Setup camera for video feed
9   async function setupCamera() {
10      const stream = await navigator.mediaDevices.getUserMedia({ video: true });
11      video.srcObject = stream;
12      await new Promise(resolve => {
13          video.onloadedmetadata = () => {
14              video.play();
15              resolve();
16          };
17      });
18  }
19
20  // 4. Emotion Prediction API
21  async function predictEmotion(blob) {
22      const form = new FormData();
23      form.append('file', blob, 'frame.png');
24      const res = await fetch('http://localhost:8000/predict', {
25          method: 'POST',
26          body: form
27      });
28      if (!res.ok) {
29          console.error('Predict API error:', await res.text());
30          return { surprise: 0, dislike: 0, attention: 0, calm: 0 };
31      }
32      return res.json();  // Return emotion data
33  }
```

Figure 5.12: Code snippet for camera setup.

The code snippet in Appendix L demonstrates the integration of the model with the webcam feed by defining the real time loop which updates the emotion every 0.2 seconds.

5.4.3 Frontend Development

The front end of the application was built using HTML, CSS, and JavaScript. The therapist was provided with a simple dashboard that displayed a video feed from the

webcam, real-time emotion predictions, and a button to start the therapy session. A visual representation of the detected emotions was also shown, updating as the session progressed as in Appendix K Video Feed Page.

The video feed from the webcam was displayed in the browser using HTML5's <video> element. Emotion predictions were sent to the frontend via Flask's SocketIO integration, enabling real-time communication between the server and the client.

5.4.4 Validation and Testing

To validate the system, the following tests were conducted:

Live Emotion Detection

The real-time emotion detection model was evaluated by comparing the predicted emotions with manual annotations of the sample expressions during therapy sessions. We validated that the frames from the live video feed were being sent to the mode; and the emotions were being classified.

Session Tracking:

The system was validated by ensuring that every session started, emotion detected, and session data were correctly stored in the database. This process started with simple user authentication. We implemented a simple Login function that took the users' credentials (user ID and password) and validated them against what was in the database.

Chapter 6 Discussion and Presentation of Findings

6.1 Introduction

This chapter interprets the evaluation results and compares the performance of the various models evaluated throughout the research. It answers key questions: *Did the models work as expected? What do the results mean?* The findings are discussed in the context of their implications for the emotion recognition system and its potential application in Dolphin-Assisted Therapy (DAT) for children with Down syndrome.

6.2 Performance of Image Based Classification Model

The baseline CNN demonstrated robust performance, particularly in classifying the Dislike category, which achieved perfect precision and recall. This suggests that the visual patterns associated with Dislike were distinct and easily learned by CNN. However, the model encountered challenges in differentiating between Attention and Calm, as evidenced by the high rate of misclassification between these two categories.

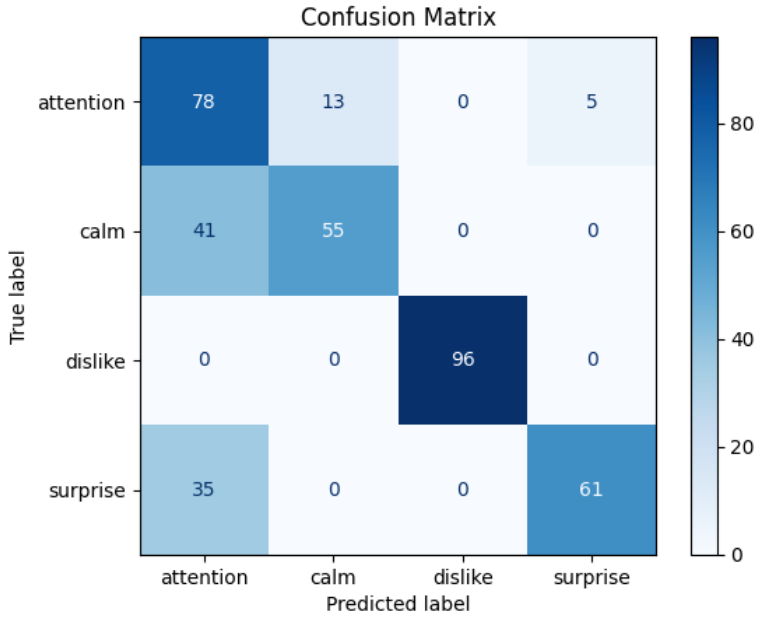


Figure 6.1 CNN Model Confusion Matrix

Specifically, many Calm instances were predicted as Attention, indicating limited sensitivity to subtle distinctions in facial expressions. Similarly, the Surprise class showed

relatively strong precision but suffered from lower recall, often being misclassified as Attention.

6.3 Performance of Feature Based Classification Model

For the feature-based selection models, similar to the CNN model, we observe distinct patterns of performance across distinct categories.

The SVM models, whether trained on AUs, Landmarks, or fused features, demonstrated challenges with the Attention category, where precision and recall were both low. This suggests that the model struggled to differentiate Attention from other emotions, particularly Calm. Conversely, the Dislike class exhibited perfect recall and high precision in all SVM configurations, indicating the clear distinction of this emotion in the feature set.

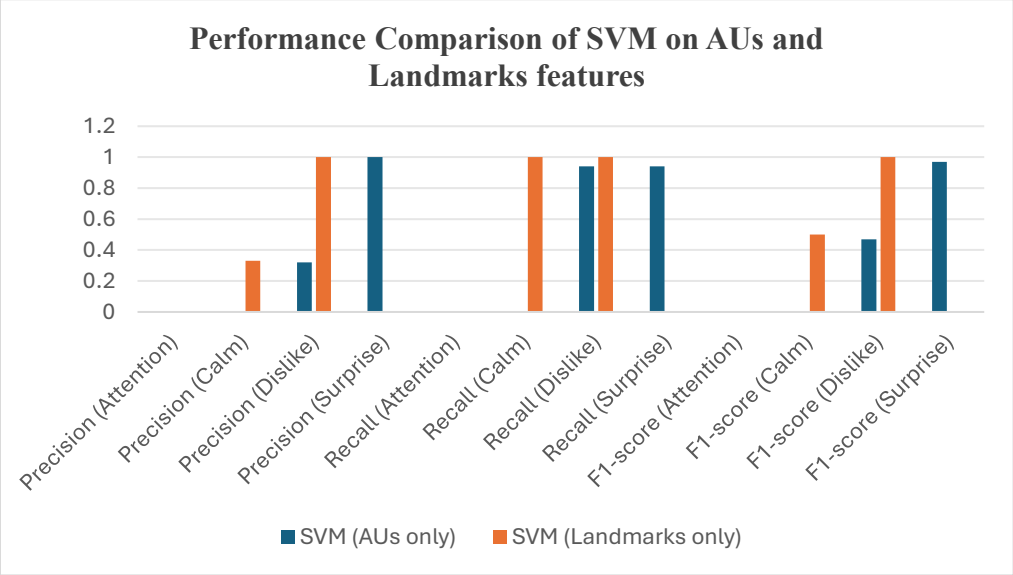


Figure 6.2 SVM Performance

In contrast, the Random Forest (RF) model showed moderate performance across the different feature sets. RF (Fused) performed slightly better than the others, especially for Dislike, where it achieved a higher F1-score. This is likely due to the ability of RF to handle complex, non-linear decision boundaries. However, it still struggled with Attention and Surprise, where misclassifications were observed.

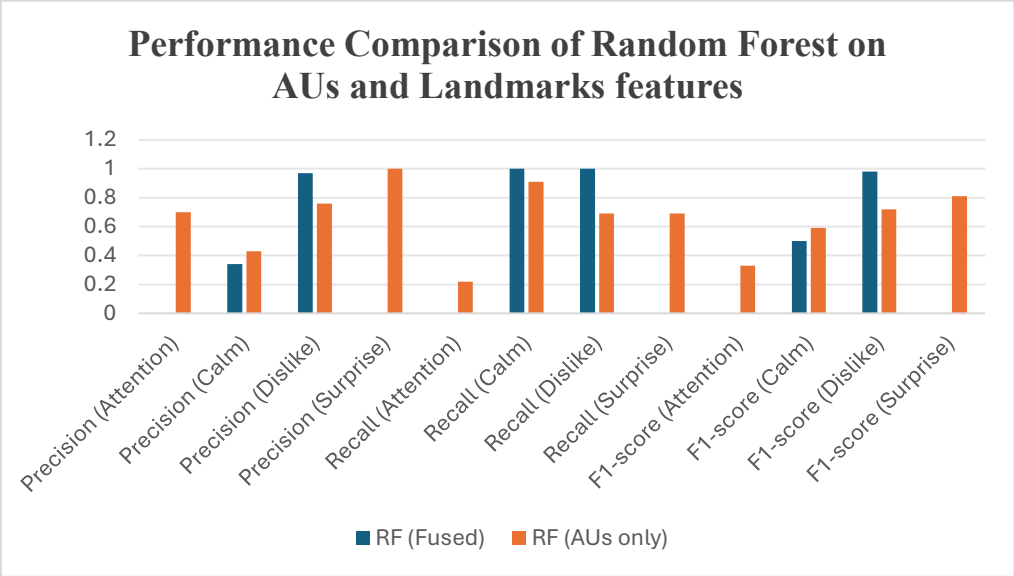


Figure 6.3 RF Performance

For MLP, the fused features led to a slight improvement in classifying Attention, but it still faced difficulties, particularly in distinguishing classifying Dislike.

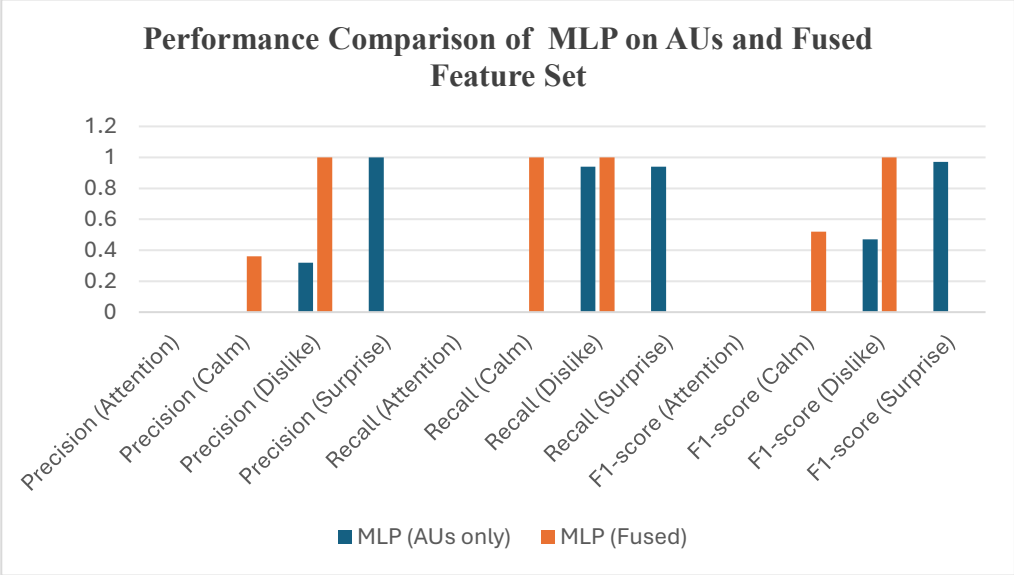


Figure 6.4 MLP Performance

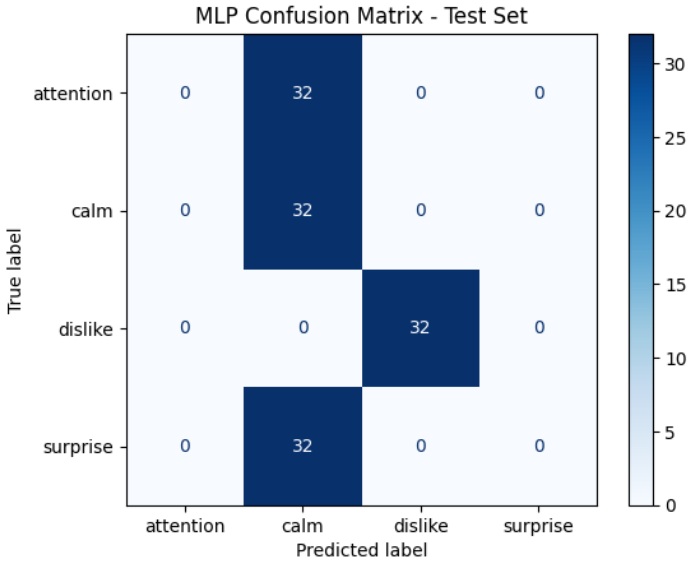


Figure 6.5 Confusion Matrix For MLP

Despite the improvements, MLP (Fused) and MLP (Landmarks) yielded similar results in the lower performance categories, with F1-scores at 0 for Attention.

The Deep Neural Networks (DNN) models did not show significant improvement over the RF or SVM models. However, DNN (Fused) had a modest gain in F1-score for Dislike and Calm.

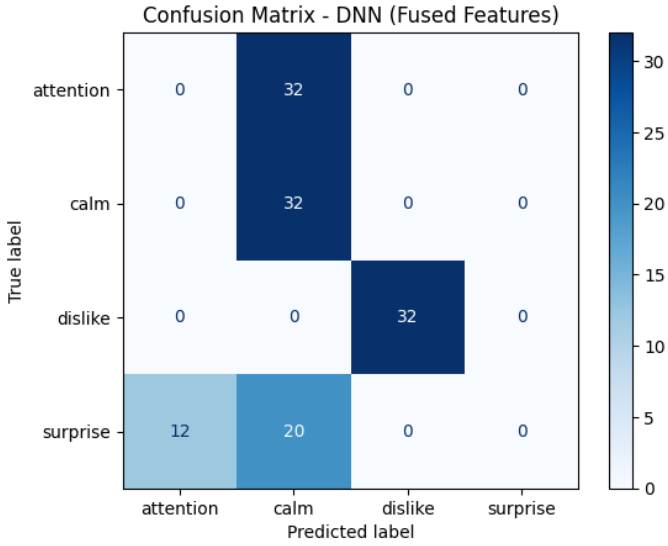


Figure 6.6 DNN Confusion Matrix

6.4 Performance of Hybrid Classification Model

For the hybrid model, we observed that the "Dislike" class had extremely high accuracy with a perfect score of 1 for both precision and recall. However, for other classes like "Attention" and "Surprise,", it struggled considerably. Some misclassifications were evident, particularly between these classes as shown below.

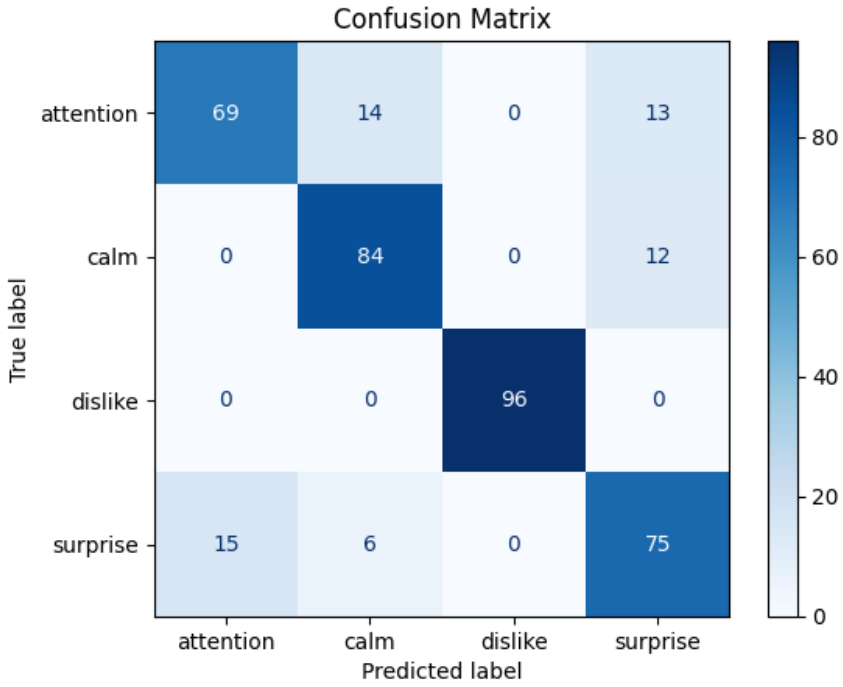


Figure 6.7 Performance of Hybrid Model

6.5Results Discussion and Comparative Analysis

The results revealed some difficulties in distinguishing the “Attention” and “Surprise” classes. To better understand the reasons behind the varying performance across different classes, several techniques were employed. PCA and t-SNE were utilized to visualize the distribution of features across the different emotion classes.

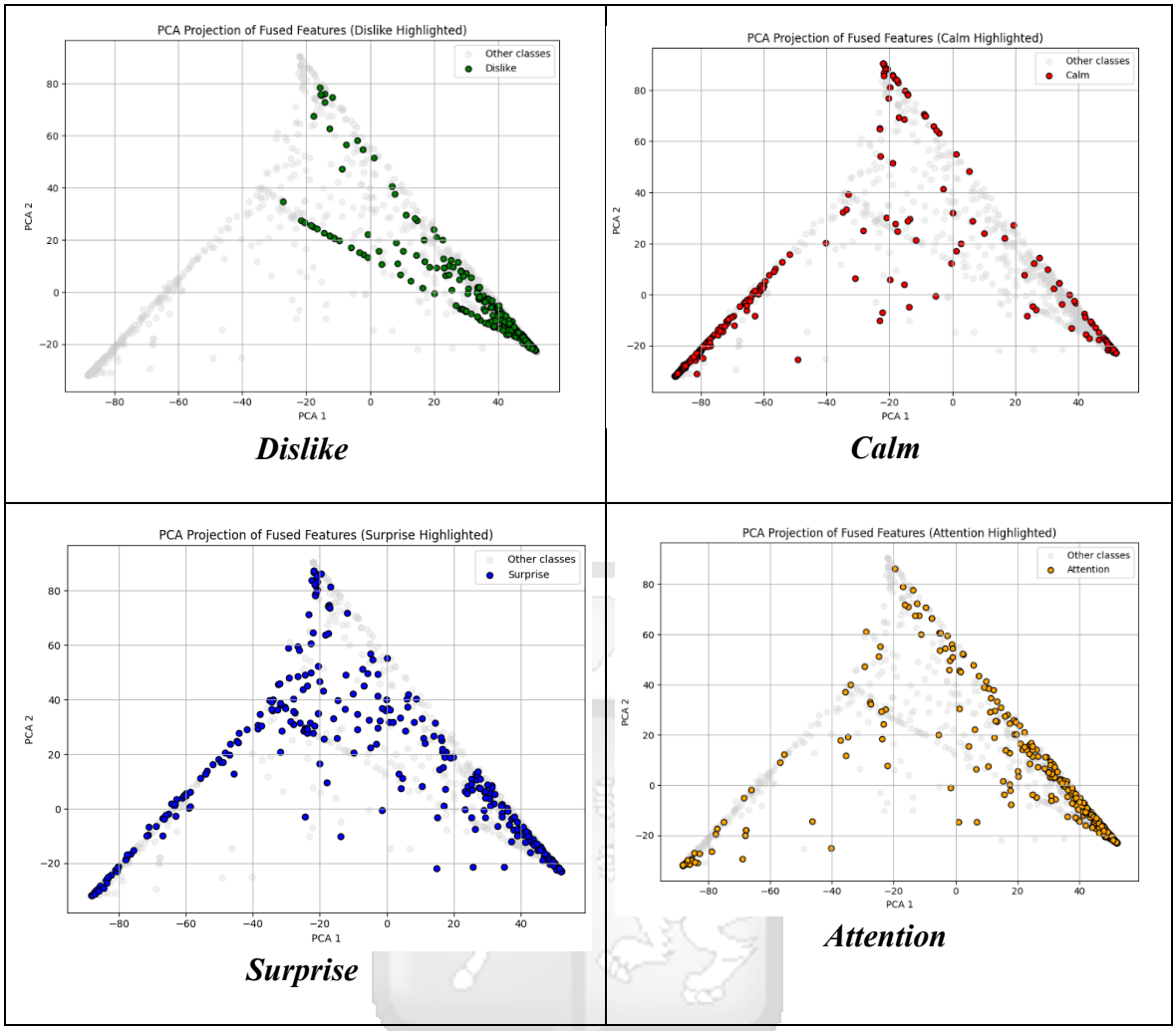


Figure 6.8 PCA Comparison across the four classes.

By observing the clusters in the plots, we could discern that classes like Calm and Dislike had a more distinct distribution compared to Attention, which showed overlapping features with other classes, leading to lower classification accuracy for Attention.

In addition to the visual techniques, we performed an ANOVA statistical test to identify which features exhibited the most variation across the classes. For instance, the features with the highest F-values, such as Feature 48, Feature 945, and Feature 39, had a marked difference across three of the four classes. These features were crucial in distinguishing between emotions like Surprise, Dislike, and Calm. However, Attention class performed poorly in comparison, with fewer significant features contributing to its classification. These

suggested that Attention lacked distinctive facial patterns that could be easily captured by the model. This observation may explain why Attention was often misclassified as other emotions like Dislike, despite their contrasting facial expressions.

The visualizations further revealed an overlap between dislike and surprise plots, suggesting that the features for these two emotions are similar or that the classifier may have difficulty distinguishing between them. This was consistent with the classification difficulties observed in some of the models.

The feature-based models, which combined both facial landmarks and action units (AUs), had the potential to offer more comprehensive insights into the emotional expressions but did not consistently outperform the other models. While the engineered dataset introduced richer features, it also introduced more complexity, and neural network struggled to learn the optimal patterns in the data. The increased dimensionality may have also led to overfitting, where the model memorized noise in the data leading to poor generalizability.

The CNN model, trained on raw image data, showed more robust performance in some areas but struggled with less visually distinct emotions such as attention and calm (a phenomenon observed with all the models. In conclusion, the difficulty of distinguishing between Attention and Calm, in particular, stemmed from the inherent similarity in their facial expressions and the challenges posed by the dataset's characteristics.

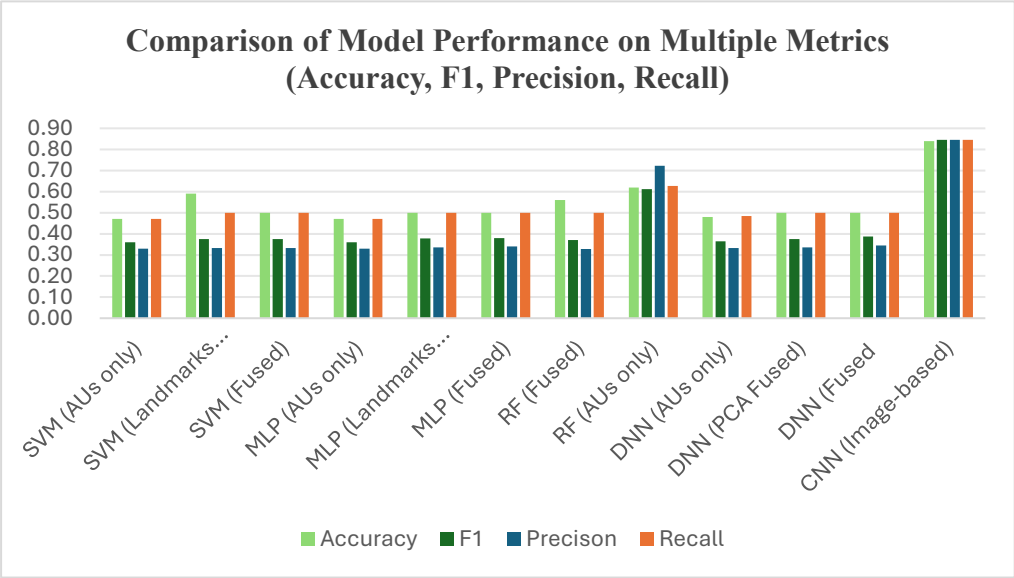


Figure 6.9 Comparison of Model Performance Metrics

Chapter 7 Conclusions and Recommendations

7.1 Conclusion

This study developed and evaluated an ML-powered emotion recognition system, specifically designed for children with Down syndrome undergoing therapy. It critically builds on existing emotion recognition research by specifically addressing a niche but underexplored population in a therapeutic setting (dolphin-assisted therapy). While many prior studies in emotion recognition focus on neurotypical adults or generalized populations, few incorporate individuals with cognitive or developmental disorders, and even fewer contextualize the analysis within assistive therapeutic environments. Moreover, most existing works rely on multimodal datasets (e.g., EEG + facial data), which, while powerful, are not always practical in sensitive or real-world clinical scenarios. Due to data limitations and ethical considerations, this study focused solely on facial imagery and facial landmarks/AUs extracted via Mediapipe and DeepFace; demonstrating that meaningful emotional assessment is still feasible with constrained modalities.

The findings highlight the challenges and potential of integrating machine learning models for emotion recognition in neurodiverse populations. The key conclusions drawn from this study are as follows:

i ML Integration Improves Emotion Recognition

The ability to recognize emotions in children with Down syndrome significantly improved through the use of deep learning models. By leveraging facial landmarks and action units, the models were able to better classify subtle emotional expressions that are typical in neurodiverse populations. However, most of the models struggled to distinguish emotions, particularly between Attention and Calm, which points to the need for mor research and better data collection

ii A Hybrid Model/ Approach Offers Enhanced Understanding of Emotional Cues

From this study we combined facial landmarks and action units. The results showed a lot of promise; by capturing a richer set of features we created a high-dimensional data set. Due to the size of the dataset, the models still suffered from overfitting. In the end, the hybrid model ultimately outperformed other unimodal models in distinguishing the four emotions, with a slightly better performance for Attention and Calm. This suggests that multi-modal data can enhance emotion recognition systems

iii Web Application Deployment for Real-time Recognition

The system was successfully implemented as a web application using the Flask API, enabling real-time emotion recognition during therapy sessions. This application allowed the therapists to receive instant feedback on the emotional state of the child, fulfilling the set objectives; to develop an adaptive emotion recognition system. The web app is an accessible option and feasible for everyday use.

iv Potential for Behavioral Insights in Therapy

By integrating emotion recognition into clinical therapy, the application can potentialy improve therapeutic outcomes by providing accurate classification of the children's emotional states. This can help in both in identifying emotional distress or positive engagement.

7.2 Limitations and Delimitations

This study ; as defined in the scope in Chapter 1 focused on emotion recognition from facial expressions in children with Down syndrome which exclude any other factors that may contribute to emotional understanding. It did not take into account the context or environment. While the choice to limit the scope allowed for a more focused investigation it would be worth considering other data sources. Another delimitation was the choice of

machine learning algorithms. While the study implemented both shallow and deep learning models, it did not explore all possible architectures or ensemble methods that could potentially improve performance, such as attention-based models or hybrid systems that integrate temporal aspects of emotion.

The limitations of this study primarily stem from the nature of the dataset and the complexity of emotion recognition in children with Down syndrome. The dataset was still relatively small, especially in comparison to larger, more diverse datasets that are normally used in emotion recognition tasks. This definitely limited the model's ability to generalize across all emotion classes.

7.3 Recommendations

In the current study, while the validation set was crucial for monitoring model performance during training, and techniques like early stopping and checkpointing were applied, no hyperparameter tuning was performed. This decision was deliberate, primarily due to the relatively small size of the dataset, even after augmentation. Applying exhaustive hyperparameter optimization methods, such as grid search or randomized search, could lead to overfitting, especially given the potential for the model to become excessively fine-tuned to the validation set. Overfitting would, in turn, compromise the generalizability of the model, which is a crucial aspect when deploying in real-world applications.

Additionally, the baseline CNN model was intentionally kept simple to establish a benchmark, with hyperparameters (e.g., number of filters, dropout rate, batch size, and learning rate) selected based on established best practices for small image classification tasks. The simplicity of the model allowed for a clearer comparison of more complex models in subsequent experiments. However, in future work, incorporating hyperparameter tuning through methods like cross-validation or Bayesian optimization will be crucial. These

methods allow for a more systematic search for optimal parameters while reducing the risk of overfitting. An independent validation strategy would be necessary to prevent data leakage and ensure a more accurate and robust model evaluation.

7.4 Future works

The findings suggest avenues for future improvements. First, expanding the dataset size, either by collecting more data could help address class imbalance and improve the model's ability to generalize well to unseen data. This can be through augmentation techniques or more advanced synthetic data generation. Furthermore, there is room for research in this area by looking into complex architecture. Transformers or hybrid models that integrate both visual and contextual features without the fusion step could offer enhanced performance. Such models have already been utilized in today's Generative AI and have already shown promise in handling dynamic and complex data. They may be able to capture subtle shifts in emotion that simpler models struggle with.

References

Agung, E. S., Rifai, A. P., & Wijayanto, T. (2024). Image-based facial emotion recognition using convolutional neural network on emognition dataset. *Scientific Reports 2024 14:1*,

Antonarakis, S. E., Skotko, B. G., Rafii, M. S., Strydom, A., Pape, S. E., Bianchi, D. W., Sherman, S. L., & Reeves, R. H. (2020). Down syndrome. *Nature Reviews Disease Primers 2020 6:1*, *6*(1), 1–20. https://doi.org/10.1038/s41572-019-0143-7

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Boato, E., Melo, G., Filho, M., Moresi, E., Lourenço, C., & Tristão, R. (2022). The Use of Virtual and Computational Technologies in the Psychomotor and Cognitive Development of Children with Down Syndrome: A Systematic Literature the Use of Virtual and Computational Technologies in the Psychomotor and Cognitive Development of Children with Down Syndrome: A Systematic Literature Review. https://doi.org/10.3390/ijerph19052955

Breuer, R., & Kimmel, R. (2017). A deep learning perspective on the origin of facial expressions.

Chen, Y., Zhang, S., & Liu, Z. (2022). Emotion-based personalized interaction in human-computer interfaces. *International Journal of Human-Computer Interaction, 38*(9), 879-890.

Cohn, J. F., De la Torre, F., & Matthews, I. (2019). The promises and perils of automated facial action coding in affective computing. *Frontiers in Neuroscience*, 13, 1–12.

Davison, A. K., Lansley, C., Ng, C. C., Tan, K., & Yap, M. H. (2016). Objective micro-facial movement detection using FACS-based regions and baseline evaluation.

Day, T. N., Mazefsky, C. A., & Wetherby, A. M. (2022). Characterizing difficulties with emotion regulation in toddlers with autism spectrum disorder. *Research in Autism Spectrum Disorders*, *96*. https://doi.org/10.1016/j.rasd.2022.101992

Ekman, P., & Friesen, W. V. (1978). *Manual for the Facial Action Coding System*. Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (2002). *Facial Action Coding System: The manual on CD ROM*. Research Nexus.

Ekman, P., Sorenson, R., & Friesen, W. (1969). Pan-Cultural-Elements-In-FacialDisplay-Of-Emotion.pdf. *SCIENCE*, *164*, 86–88.

Esbensen, A. J., Schworer, E. K., & Hartley, S. L. (2024). Down Syndrome. *Contemporary Clinical Neuroscience*, *Part F3408*, 279–302. https://doi.org/10.1007/978-3-031-66932-3_13

Global Down Syndrome Foundation – ARDownSyndrome.org. (n.d.). Retrieved October 15, 2024, from https://ardownsyndrome.org/global-down-syndrome-foundation/

Guo, R., Guo, H., Wang, L., Chen, M., Yang, D., & Li, B. (2024). Development and application of emotion recognition technology — a systematic literature review. In *BMC Psychology* (Vol. 12, Issue 1). BioMed Central Ltd.https://doi.org/10.1186/s40359-024-01581-4

Jacintha, V., Simon, J., Tamilarasu, S., Thamizhmani, R., Thanga Yogesh, K., & Nagarajan, J. (2019). A review on facial emotion recognition techniques. *Proceedings of the 2019 IEEE International Conference on Communication and Signal Processing, ICCSP 2019*, 517–521. https://doi.org/10.1109/ICCSP.2019.8698067

Jones, M., & Smith, A. (2023). Emotion detection systems in educational settings: Enhancing learning outcomes. *Journal of Educational Technology, 45*(2), 115-127.

Kamiri, J., & Mariga, G. (2021a). Research Methods in Machine Learning: A Content Analysis. International Journal of Computer and Information Technology (2279-*0764)*, *10*(2). https://doi.org/10.24203/ijcit.v10i2.79

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annu. Rev. Psychol.*, *66*, 799–823. https://doi.org/10.1146/annurev-psych-010213-

Lien, J. J., Jenn, D. H., & James, J. (1998). Automated facial expression recognition based on FACS action units. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 200–205.

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. In *Electronics (Switzerland)* (Vol. 10, Issue 10). MDPI AG. https://doi.org/10.3390/electronics10101163

Liu, Y., Wang, L., & Deng, W. (2020). A systematic review of facial action coding system in human emotional behavior research. *Frontiers in Psychology*, 11, 920.

Mahalakshmi, G. (2017). Emotion Models: A Review. In *Article in International Joural of* Control Theory and Applicationshttps://www.researchgate.net/publication/319173333

Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology*. MIT Press.

Moreno Escobar, J. J., Morales Matamoros, O., Aguilar del Villar, E. Y., Quintana Espinosa, H., & Chanona Hernández, L. (2023). DS-CNN: Deep Convolutional Neural Networks for Facial Emotion Detection in Children with Down Syndrome during Dolphin-Assisted Therapy. *Healthcare (Switzerland)*, *11*(16). https://doi.org/10.3390/healthcare11162295

Ouzar, Y., Bousefsaf, F., Djeldjli, D., & Maaoui, C. (n.d.). Video-based multimodal spontaneous emotion recognition using facial expressions and physiological signals.

Oviatt, S., & Cohen, P. R. (2015). Theoretical foundations of multimodal interfaces. In *The paradigm shift to multimodality in contemporary computer interfaces* (pp. 41–52). Springer. https://doi.org/10.1007/978-3-031-02213-5_5

Pan, J., Fang, W., Zhang, Z., Chen, B., Zhang, Z., & Wang, S. (2024). Multimodal Emotion Recognition Based on Facial Expressions, Speech, and EEG. *IEEE Open* Journal of Engineering in Medicine and Biology, 5, 396.https://doi.org/10.1109/OJEMB.2023.3240280

Paredes, N., Caicedo-Bravo, E. F., Bacca, B., & Olmedo, G. (2022). Emotion Recognition of Down Syndrome People Based on the Evaluation of Artificial Intelligence and Statistical Analysis Methods. *Symmetry*, *14*(12). https://doi.org/10.3390/sym14122492

Paredes, N., Caicedo-Bravo, E., & Bacca, B. (2023). Emotion recognition in individuals with Down syndrome: A convolutional neural network-based algorithm proposal. *Symmetry*, 15(7), 1435. https://doi.org/10.3390/sym15071435

Paredes, N., Caicedo-Bravo, E., & Bacca, B. (2023). Emotion Recognition in Individuals with Down Syndrome: A Convolutional Neural Network-Based Algorithm Proposal. *Symmetry*, *15*(7). https://doi.org/10.3390/sym15071435

Pawłowski, M., Wróblewska, A., & Sysko-Romańczuk, S. (2023). Effective Techniques for Multimodal Data Fusion: A Comparative Analysis. *Sensors*, *23*(5). https://doi.org/10.3390/s23052381

Plutchik, R. (1980). A general psych evolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research, and experience. Vol. 1. Theories of emotion* (pp. 3–33). Academic Press.

Rodrigo-Claverol, M., Malla-Clua, B., Marquilles-Bonet, C., Sol, J., Jové-Naval, J., SolePujol, M., & Ortega-Bravo, M. (2020). Animal-assisted therapy improves communication and mobility among institutionalized people with cognitive impairment. *International Journal of Environmental Research and Public Health*, *17*(16), 1–14. https://doi.org/10.3390/ijerph17165899

Sahu, G., & Vechtomova, O. (2019). Adaptive fusion techniques for multimodal data.

Sideropoulos, V., Sokhn, N., Palikara, O., Van Herwegen, J., & Samson, A. C. (2023). Anxiety, concerns and emotion regulation in individuals with Williams syndrome and Down syndrome during the COVID-19 outbreak: a global study. *Scientific Reports*, *13*(1), 8177. https://doi.org/10.1038/s41598-023-35176-7

Tanabe, H., Shiraishi, T., Sato, H., Nihei, M., Inoue, T., & Kuwabara, C. (2024). A concept for emotion recognition systems for children with profound intellectual and multiple disabilities based on artificial intelligence using physiological and motion signals. *Disability and Rehabilitation. Assistive Technology*, *19*(4), 1319–1326. https://doi.org/10.1080/17483107.2023.2170478

Tanabe, S., Kato, T., & Suzuki, M. (2024). Emotion recognition system for profound intellectual disabilities using motion and heartbeat intervals. *Journal of Disability and Rehabilitation, 46*(8), 928-938.

Vazquez-Rodriguez, J. (2021). Using Multimodal Transformers in Affective Computing. 2021 9th International Conference on Affective Computing and Intelligent

Interaction Workshops and Demos, ACIIW 2021.https://doi.org/10.1109/ACIIW52867.2021.9666396

Verma, D., Kumari Barnwal, S., Barve, A., Jayanthi Kannan, M. K., Gupta, R., Swaminathan, R., & Professor, A. (n.d.). Multimodal Sentiment Sensing and Emotion Recognition Based on Cognitive Computing Using Hidden Markov Model with Extreme Learning Machine. *International Journal of Communication Networks and Information Security.* https://ijcnis.org

Yang, J., Xue, Y., Zeng, Z., & Guo, W. (n.d.). Research on Multimodal Affective Computing Oriented to Online Collaborative Learning.

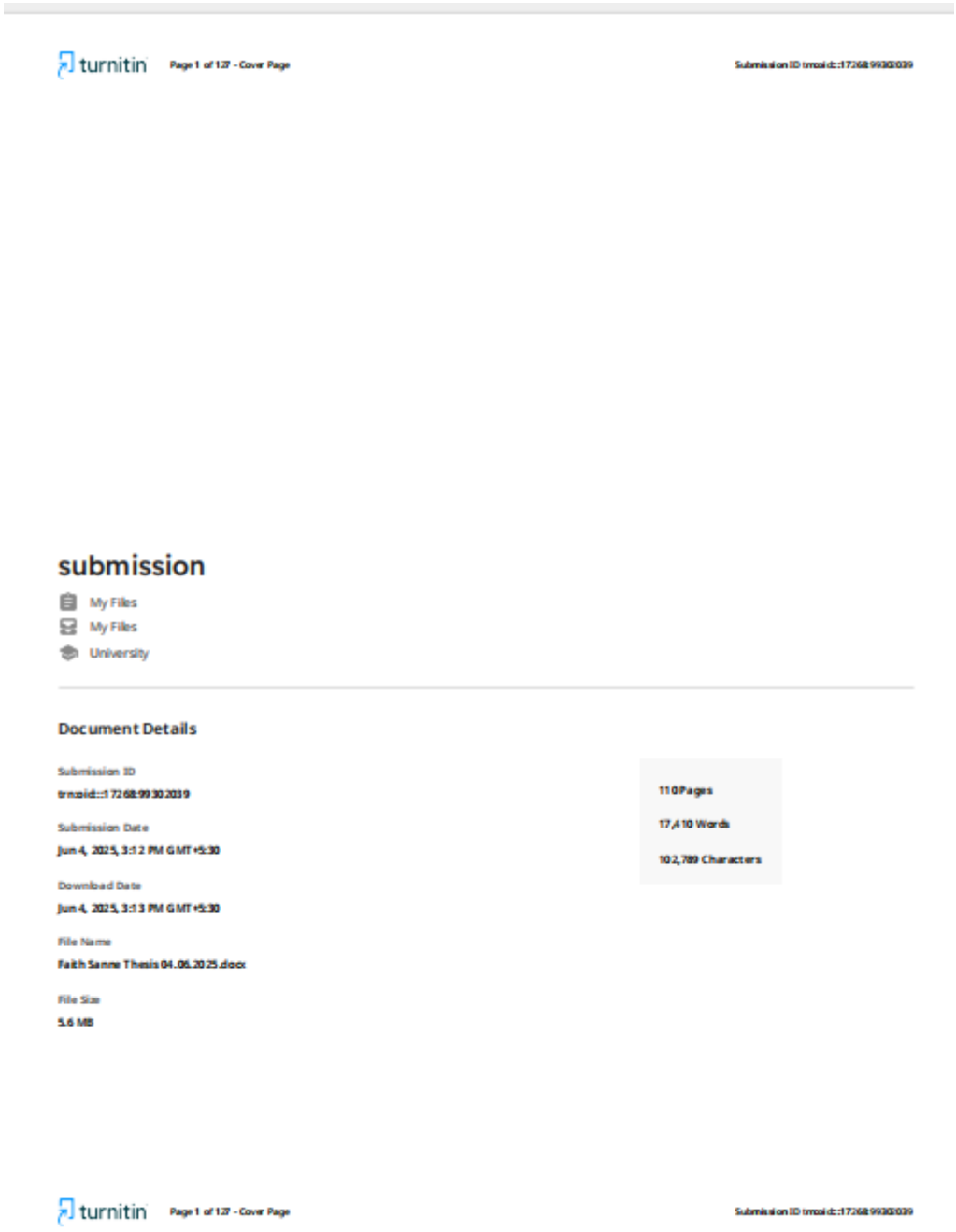
Zambrano, B. E., Gómez, B. D. L., & Morfín, V. H. C. (2021). Non-Parametric Evaluation Methods of the Brain Activity of a Bottlenose Dolphin during Assisted Therapy. *Animals: An Open Access Journal from MDPI*, *11*(2), 1–26. https://doi.org/10.3390/ANI11020417

Zhang, C. (2020). Performance Comparison Between Early Fusion and Late Fusion in Multimodal Data Processing. *Medium*.

Zhang, Y., Zhang, C., & Li, Z. (2018). A survey of affective computing: Emotion recognition and emotion synthesis. *Journal of Computer Science and Technology*, 33(4), 721–742.

Appendixes

Appendix A Plagiarism Report Page 1



turnitin Page 1 of 127 - Cover Page Submission ID trnoid::1726899302039

submission

My Files

My Files

University

Document Details

Submission ID

trnoid::1726899302039

Submission Date

Jun 4, 2025, 3:12 PM GMT+5:30

Download Date

Jun 4, 2025, 3:13 PM GMT+5:30

File Name

Faith Sanne Thesis 04.06.2025.docx

File Size

5.6 MB

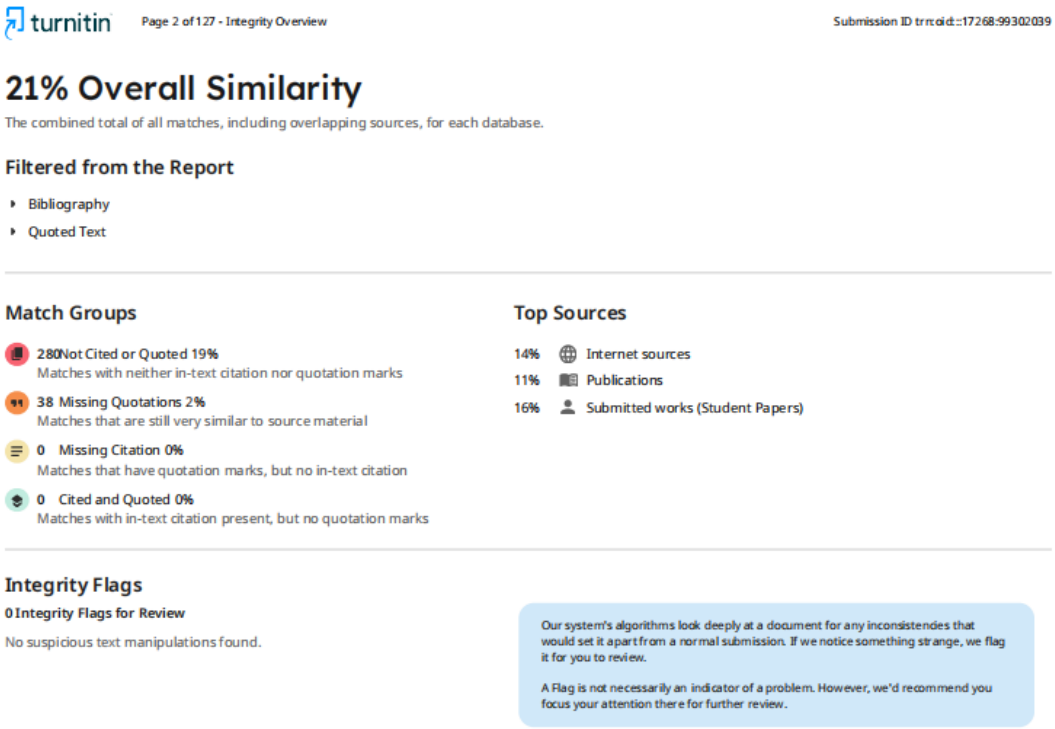
110 Pages

17,410 Words

102,789 Characters

turnitin Page 1 of 127 - Cover Page Submission ID trnoid::1726899302039

Appendix B Plagiarism Report Page 2



turnitin Page 2 of 127 - Integrity Overview Submission ID trn:oid:::17268:99302039

21% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

- 280 Not Cited or Quoted 19%
  Matches with neither in-text citation nor quotation marks
- 38 Missing Quotations 2%
  Matches that are still very similar to source material
- 0 Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

14% Internet sources
11% Publications
16% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix C Ethical Review Submission

Completion of Online Research Ethics Review Submission

You have successfully submitted your application for ethics review "USING DEEP LEARNING FOR EMOTIONAL ASSESSMENT DURING THERAPY"

Certificate awarded to: Odhiambo, Faith

Reference number: SU-ISERC2870/25

Date and Time: 2025-04-02 08:17:59

9th April 2025

Faith Sanne Odhiambo

121058

odhiambo.faith@strathmore.edu

Dear Faith,

RE: Emotional Assessment in Children with Downs Syndrome Using Deep Learning During Dolphin Assisted Therapy

This is to inform you that the Office of Graduate Studies on 9th April 2025 received your acknowledgement of breach in ethical processes given that you have already collected/analysed data and proceeded to write your Dissertation/Thesis prior to obtaining Ethical clearance. Consequently, it was noted that The Strathmore University Institutional Scientific and Ethical Review Committee (SU-ISERC) cannot review your study since you have already collected data and written the Thesis. The scientific & ethical review/approval process is ONLY done before the commencement of any experiments, implementation or any collection of data (primary or secondary-including desktop review).

This is a letter for you to proceed with the next steps of your academic requirements.

Please be advised, that in future, all research proposals should be submitted to the SU-ISERC through the RHInnO Ethics platform: https://strathmoreuniversity.rhinno.net/login

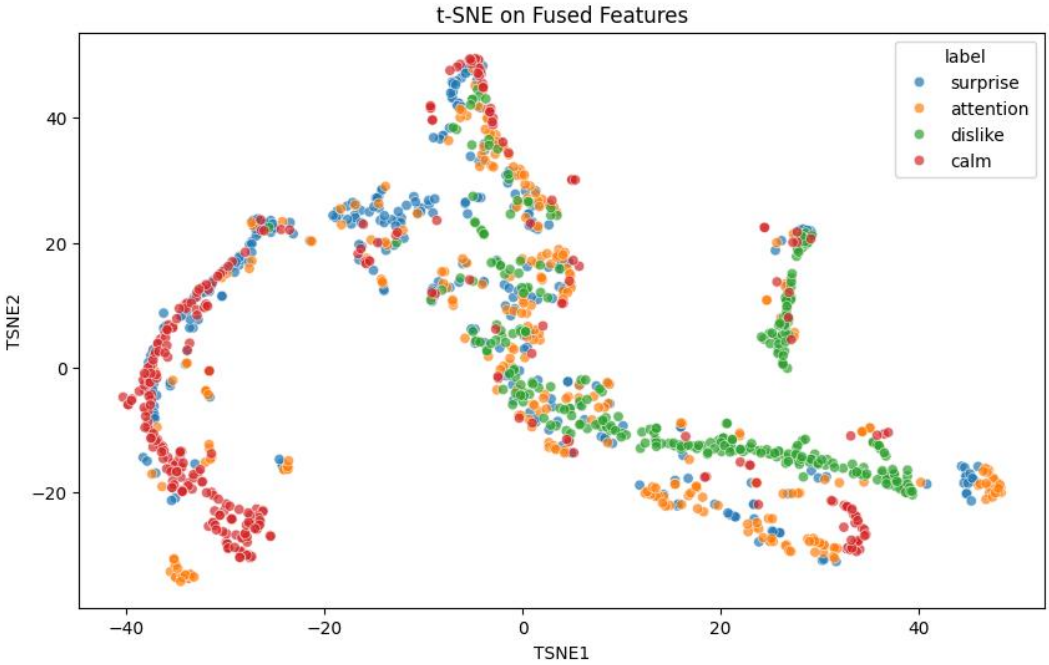
Disclaimer: *1) This is not in any way an ethical approval letter. 2) Should there be any legal implications/actions emanating from the research in terms of any ethical violations, you will be personally liable.*

Yours sincerely,

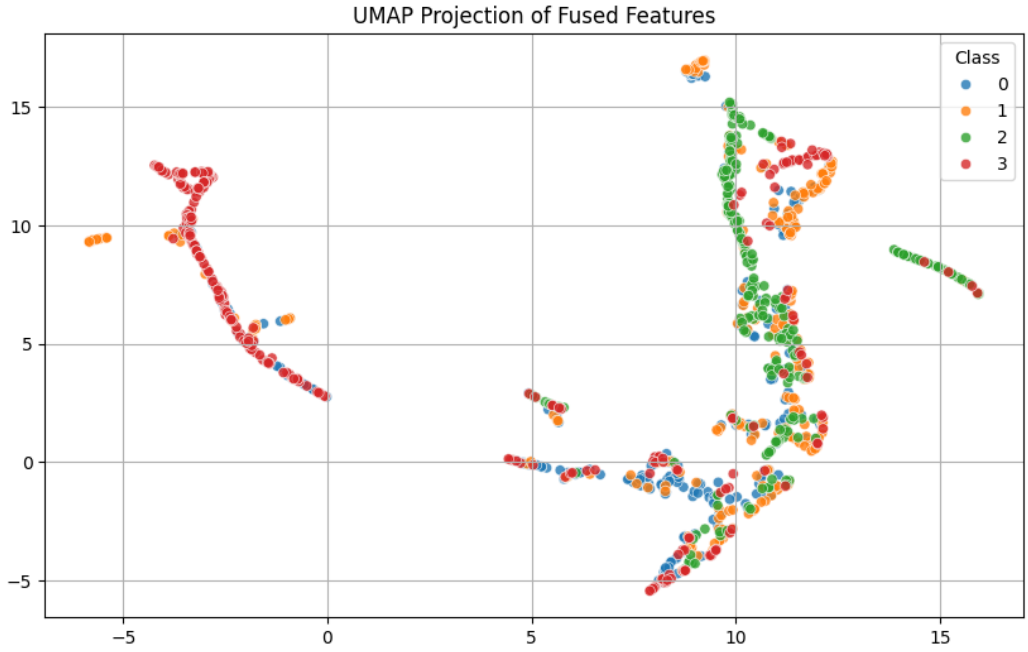
Prof. Bernard Shibwabo
Director of Graduate Studies

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix D T-SNE Visualization of Feature Set



Appendix E UMAP Projection of Fused Feature Set

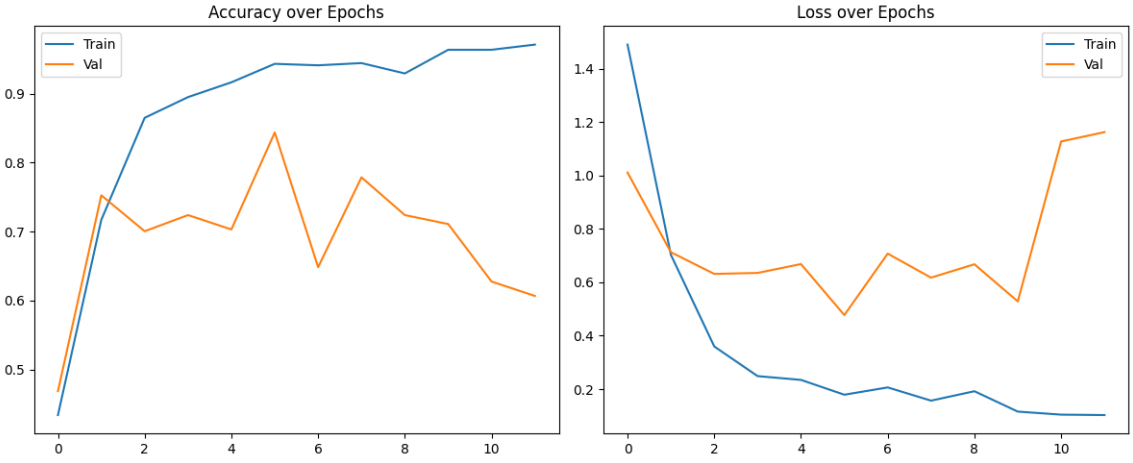


Appendix F Code Snippet for Extracting Landmarks

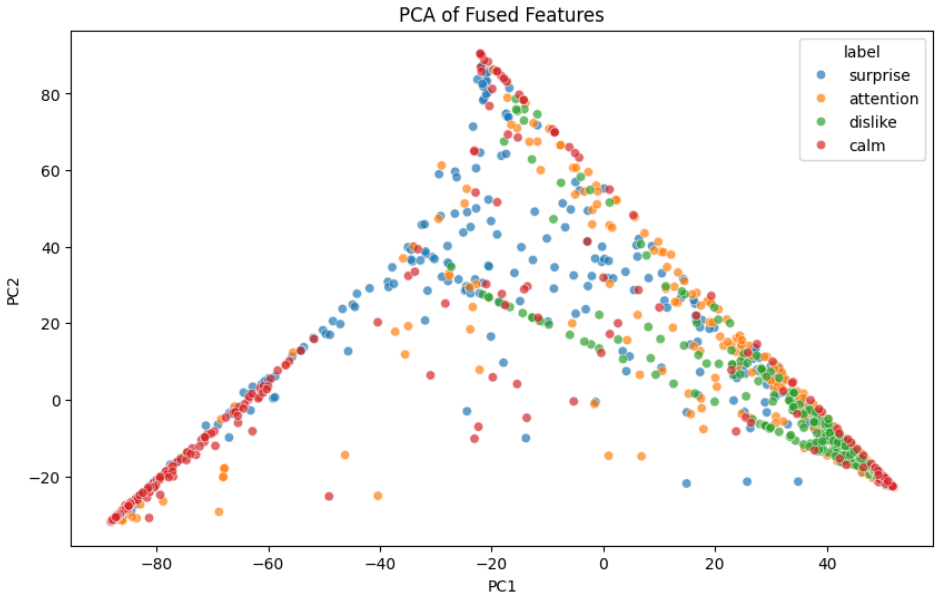
```
face_mesh = mp.solutions.face_mesh.FaceMesh(
    static_image_mode=True, # For single images
    max_num_faces=1,       # Detect only one face
    refine_landmarks=True,  # For better landmark accuracy
    min_detection_confidence=0.5)

def extract_landmarks(image_path):
    img = cv2.imread(image_path)
    rgb = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)
    results = face_mesh.process(rgb)
    if results.multi_face_landmarks:
        landmarks = results.multi_face_landmarks[0].landmark
        if len(landmarks) >= 468:
            trimmed = landmarks[:468]
            coords = np.array([[lm.x, lm.y, lm.z] for lm in trimmed])
            return coords.flatten()
        else:
            # Pad if fewer than 468
            coords = np.array([[lm.x, lm.y, lm.z] for lm in landmarks])
            pad = np.zeros((468 - len(landmarks), 3))
            full_coords = np.vstack([coords, pad])
            return full_coords.flatten()
    return np.zeros(468 * 3)
```

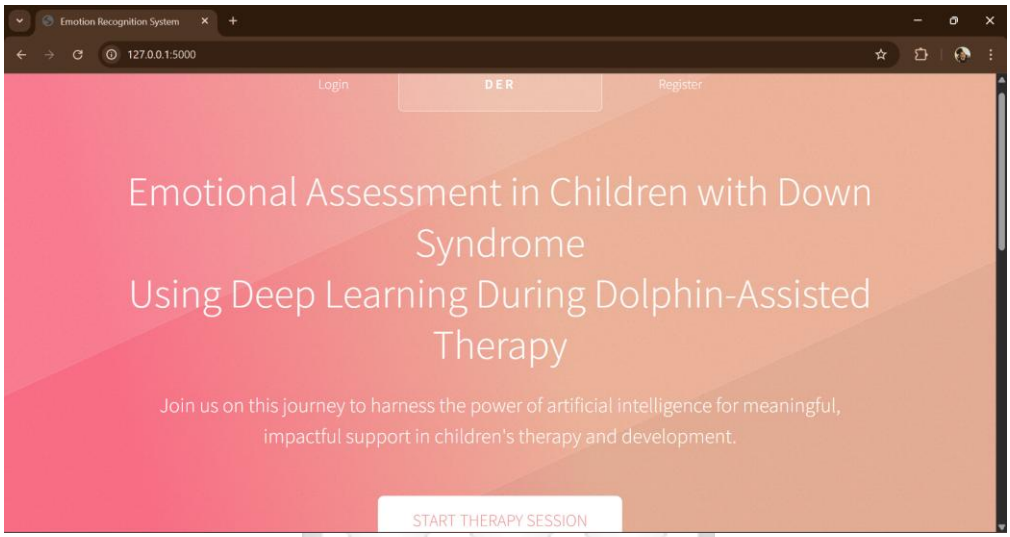
Appendix G CNN Training and Validation Performance



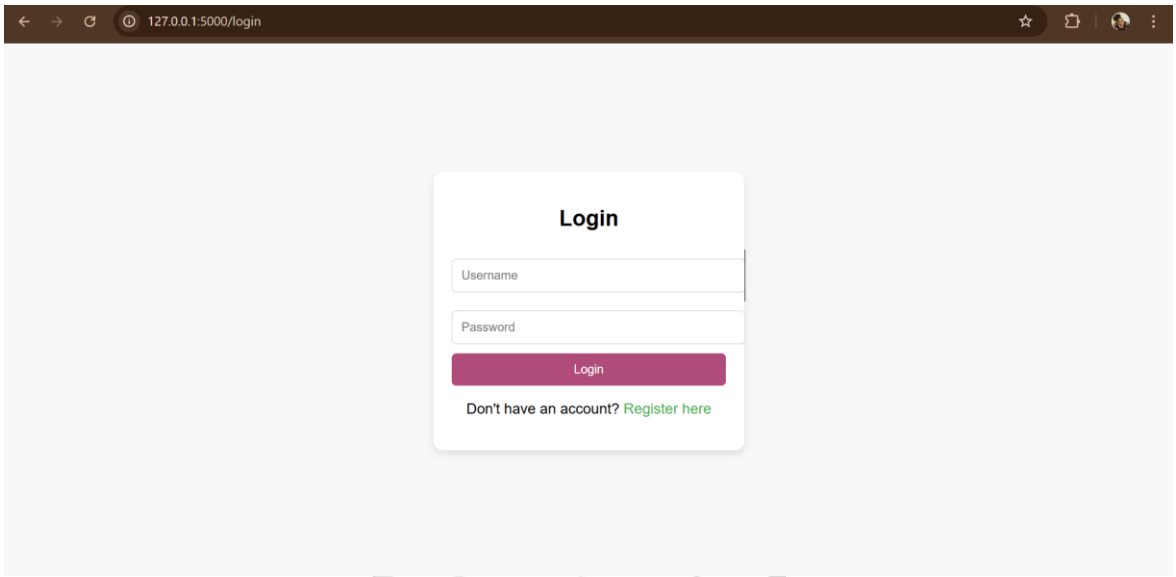
Appendix H PCA feature Map for Landmark Features



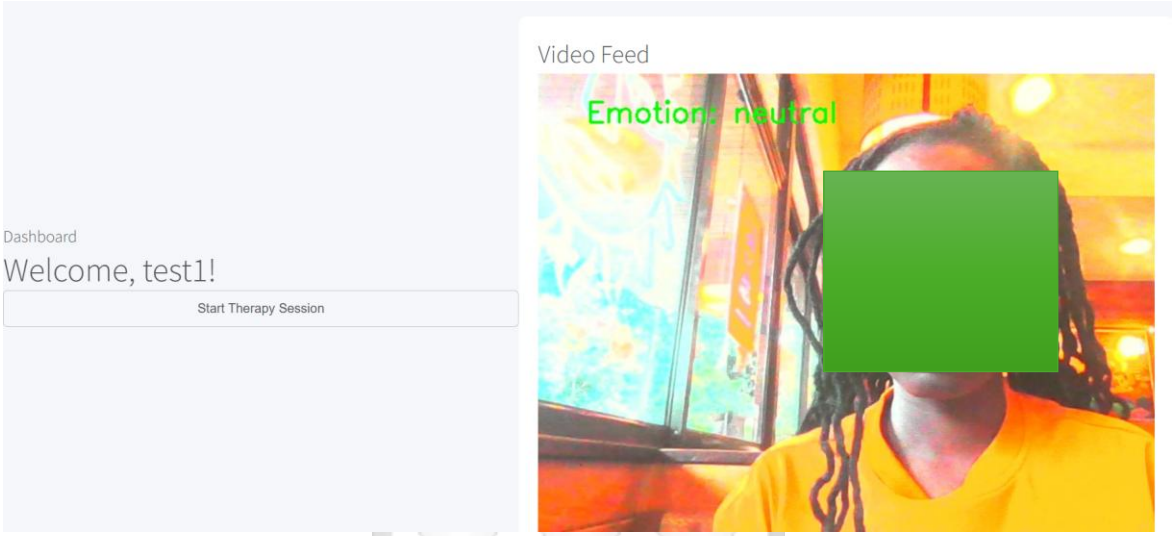
Appendix I Web Application Landing Page



Appendix J User Authentication Page(Login)



Appendix K Video Feed Page



Appendix L Code for Real Time Prediction Looping

```
35  // 5. Real-time prediction loop
36  async function detectLoop() {
37      const off = document.createElement('canvas');
38      off.width = off.height = 150;
39      off.getContext('2d').drawImage(video, 0, 0, 150, 150);
40      const blob = await new Promise(r => off.toBlob(r, 'image/png'));
41
42      // Get predictions
43      const { surprise, dislike, attention, calm } = await predictEmotion(blob);
44
45      // Update charts with real-time emotion predictions
46      updateChart(surprise, dislike, attention, calm);
47      setTimeout(detectLoop, 200);  // Call every 0.2 seconds
48  }
```