

MEDNET – An Approach to Facial Micro-Emotion Recognition using Pixel Binning and Local Binary Pattern - Convolutional Neural Network

by

Tashreef Abdullah Araf
21366023

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

Department of Computer Science and Engineering
Brac University
September 2023

© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

A handwritten signature in black ink, appearing to read 'Tashreef', is displayed within a light gray rectangular box. The signature is written in a cursive, flowing style.

Tashreef Abdullah Araf
21366023

Approval

The thesis titled “MEDNET – An Approach to Facial Micro-Emotion Recognition using Pixel Binning and Local Binary Pattern - Convolutional Neural Network” submitted by

1. Tashreef Abdullah Araf (21366023)

of Fall 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on September 2023.

Examining Committee:

External Examiner:
(Member)



Prof. Mohammad Zahidur Rahman, Ph.D.
Professor
Department of Computer Science and Engineering
Jahangirnagar University, Savar, Dhaka.

Internal Examiner:
(Member)



Dr. Amitabha Chakrabarty
Professor
Department of Computer Science and Engineering
BRAC University

Internal Examiner:
(Member)

Dr. Md. Ashrafal Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)



Dr. Amitabha Chakrabarty
Professor
Department of Computer Science and Engineering
BRAC University

Chairperson:
(Member)

Sadia Hamid Kazi, Ph.D.
Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement (Optional)

Hereby, I / Tashreef Abdullah Araf / consciously assure that for the manuscript “MEDNET – An Approach to Facial Micro-Emotion Recognition using Pixel Binning and Local Binary Pattern - Convolutional Neural Network” the following is fulfilled:

- 1) This source material is the authors’ own original work, which has not been previously published elsewhere.
- 2) The paper reflects the authors’ own research and analysis in a complete and truthful manner
- 3) The paper properly credits the meaningful contributions of relevant research.
- 4) The results are appropriately placed in the context of prior and existing research.
- 5) All sources used are properly disclosed. Literally copying of text must be indicated as such by using quotation marks and giving proper reference.

The violation of the Ethical Statement rules may result in severe consequences.

A handwritten signature in black ink, appearing to read 'Tashreef', is centered within a light gray rectangular box. The signature is fluid and cursive.

Tashreef Abdullah Araf
21366023

Abstract

Facial-Expression recognition is a very intriguing field of research, due to the complexity in its approach and applicability of widely available databases. However, Micro-expression recognition is quite a vague yet growing area of research due to its applicability in revealing minute facial expressions. These emotional triggers happen only under very pressing circumstances, which means detecting them can also be extremely tough due to shortage of time during which it lasts. In this study, the approach to Micro-facial expression detection is to explore passive and real-time observation that produces a great result for micro-facial expression recognition using a vast data set trained using new training techniques. A total of 59 papers were analyzed whose concepts were associative to our main thesis concept, which were categorized into three stages: Construction of a new dataset which constituted of standard and new facial images, which was trained using innovative image processing pipelines, implementation of a new Binary Pattern layer our Neural Network layer to accelerate the models expression tracking abilities, creation of a new facial model capable of facial and micro-facial expression recognition that performs better statistically when compared to its counterparts. Furthermore, the new model was tested in both artificial and real-world scenarios to accentuate the reliability of the data sources.

Keywords: VisageEmotioNet, Facial Expression, Micro-Facial Expression, MED-Net, OpenCV, Pixel Binning, LBP

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, I am grateful to my supportive family, without the help of whom none of this would have been possible. I am thankful for the mental and financial support of my parents in this journey. Thirdly, to my advisor, Dr Golam Rabiul Alam sir, for his kind support and advice in our work. He helped us whenever we needed guidance

Finally, I thank the internal and external examiners who have given me a platform to showcase my hard-work.

Table of Contents

Declaration	i
Approval	ii
Ethics Statement	iv
Abstract	v
Acknowledgment	vi
Table of Contents	vii
List of Figures	ix
List of Tables	x
Nomenclature	xi
1 Introduction	1
1.1 Introduction	1
1.2 Motivation	1
1.3 Research Contribution	3
1.4 Thesis Organization	3
2 Related Work	4
3 Methodologies	13
3.1 Top Level View of MEDNet	13
3.2 Data Collection	15
3.3 Overall System Model with Description	17
3.4 Dataset Training Approach	17
3.5 Data Preparation	19
3.6 Local Binary Pattern	19
3.7 Proposed Pixel Binning Approach	21
3.8 Background Separation	23
3.9 Mediapipe Facial Movement Detection	25
3.10 Haarcascade Classifier implementation	27

4	Analysis of Result	29
4.1	Performance Evaluation Metric	29
4.2	Comparative Analysis with VGGface and FaceNet	32
4.2.1	Expression Recognition	32
4.2.2	Facial Non-Appearance Recognition	33
4.2.3	Face and Name Recognition	35
4.2.4	Expression Recognition Detection Speed	36
4.3	Test System Analysis	37
4.4	Performance Analysis in Real-Time and Video	37
5	Conclusion	41
	References	42

List of Figures

1.1	The valence–arousal continuous emotional space	2
2.1	Fer2013 Dataset	8
2.2	RAFDB Dataset	9
2.3	FEER Dataset	10
2.4	AffectNET Dataset	11
2.5	VisageEmotioNet Dataset	11
2.6	Categorization of Macro and micro Expression Datasets	12
3.1	Top Level View of Proposed MEDNet Architecture	14
3.2	Image processing pipeline through Background Seperation and LBP	17
3.3	CNN	19
3.4	9x1 Pixel Binning	22
3.5	Dataset accuracy before and after Proposed Pixel Binning Approach .	22
3.6	Background Separation Flowchart	24
3.7	Erosion	25
3.8	Dilation	25
3.9	Mediapipe Implementation	26
3.10	MEDnet Classification using Mediapipe	26
3.11	Haarcascade Classifier	28
4.1	Confusion Matrix comparison between various datasets	29
4.2	Normalized Confusion Matrix of MEDNet Model with VisageEmo- tioNet Dataset	30
4.3	Accuracy per epoch	30
4.4	Loss per epoch	31
4.5	VisageEmotioNet Performance Parameters	31
4.6	FEER Performance Parameters	31
4.7	Fer2013 Performance Parameters	32
4.8	RAF-DB Performance Parameters	32
4.9	Expression Recognition Comparison	33
4.10	Face and Non-Face Recognition for models	34
4.11	Face to Non-Face Ratio	35
4.12	Comparison of Models in terms Facial and Nomenclature Data	36
4.13	Average and Maximum Frametime	37
4.14	MEDNet Live Feed	40
4.15	In the wild Mediapipe Implementation	40

List of Tables

3.1	Number of Images per Expression (Fer2013)	15
3.2	Number of Images per Expression (RAF-DB)	15
3.3	Number of Images per Expression (FEER-DB)	16
3.4	Number of Images per Expression (VisageEmotioNet)	16

List of Symbols

The next list describes several symbols that will be later used within the body of the document

BSKNN Background Separation K-Nearest Neighbour

CNN Convolutional Neural Network

DT Decision Tree

KNN K-Nearest Neighbour

LBP Local Binary Pattern

MEDNet Micro-Expression Detection Network

PB Pixel Binning

RGB Red-Green-Blue

VGGFace Visual Geometry Group Face

HOG Histogram of Oriented Gradients

Chapter 1

Introduction

1.1 Introduction

Facial Expression [16] has always been a very intriguing field of study and has found very good success in overall feature extraction and classification. One main reason for this is FER, research dedicated to Facial Expression classification, which is a direct response to advances in field of medical science, Marketing, Social Media Research, Robot AI automation [4] etc. However, most of the research dedicated for this purpose has always been to find Macro-facial expressions and most of the dataset training has been focused mostly on that. But unfortunately, the traditional approach does not really allow for Micro-expression training, which requires precision to a much more minute time constraint. Wu et al. ; Liong et al. and Liong et al. [13] introduced a model and evaluated the performance of their models using apex micro-expression images and static feature extraction techniques with subpar recognition results. The approach that have been made here are modified versions of a lot of previous techniques used. MEDNet uses a LBP-CNN structuring where the the LBP was used during the training stage to prepare the dataset using Background Subtraction algorithm and CNN was used to classify the image data base that has been extracted and differentiated using CNN. All of the images used to create the model was trained through an extensive Image Processing pipeline which includes a Pixel-binning Algorithm.

1.2 Motivation

The first successful attempt at Facial Expression Research was done by US Transportation Security Administration (TSA) in the 1960s. The primary source of consultation was Paul Ekman[3], Professor of psychology at the University of California, San Francisco. Professor Ekman had developed a method to identify minute facial expressions and map them on to corresponding emotions. This method was called “behavior detection officers” which scanned faces to find anomalies in detection in the faces of the officers. But prior to this, USTSA wanted to consult Ekman due his earlier attempts at a labelling method facial recognition two decades earlier which Paul Ekman developed in association with Wallace V Friesen called “Emotion Facial Action Coding System (EMFACS)”, which is considered principally to be the gold standard for measuring emotion during the 1980s. Later on, an extensive recognition dataset called fer2013 was introduced which allowed better recognition. This showed

that software-based training and testing can be done and be improved upon extensively. In terms of Facial Expression, the four quadrants of Valence and Arousal can be applied:

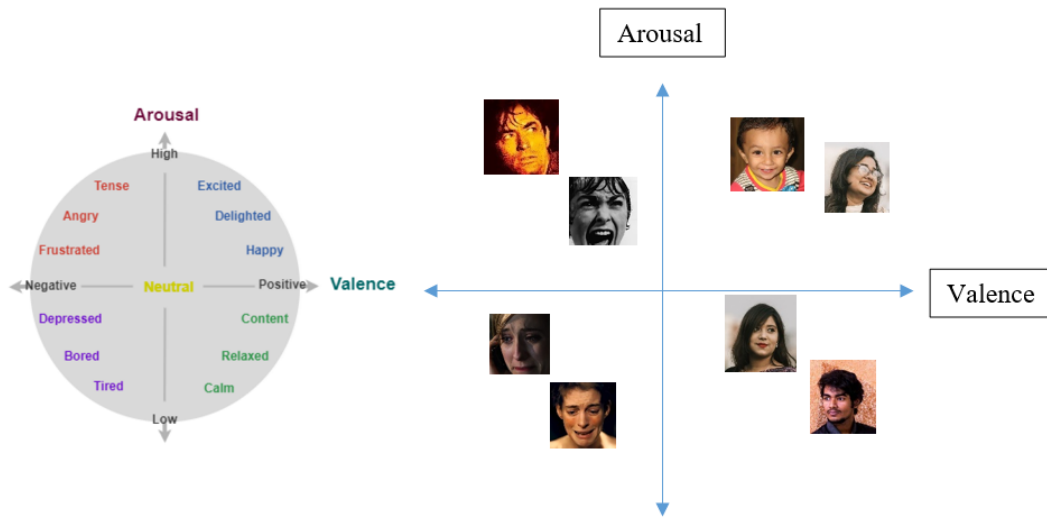


Figure 1.1: The valence–arousal continuous emotional space

First quadrant—emotional states go from pleased (high valence, medium arousal) to excited (about neutral valence, high arousal);
 Second quadrant—high arousal with about neutral valence here indicates an alarmed state, while high-negative valence and medium arousal bring to a frustrated state;
 Third quadrant—in this quadrant, high-negative valence and medium arousal indicate sad/depressed condition, while the status with arousal and about neutral valence corresponds to a tired state;
 Fourth quadrant—finally, in this quadrant arousal and about neutral valence a calm/sleepy state is valence and medium arousal.

In terms of Micro-expression recognition[1], one trend is represented by the shift from posed to spontaneous and in-the-wild capturing conditions. In particular: - Posed datasets are typically acquired by asking the subjects to show one of the six basic expressions as defined by Paul Ekman himself. In most of the cases, actors with really defined facial-expressions are enrolled, and capturing takes place in restricted conditions. - Spontaneous datasets include expressions that are given in by the participants. For example, the results may be extrapolated from watching a video or a face-to-face interaction. Participants are aware that they are monitored, but the emotions are always natural. - In-the-wild datasets are done in the real-world scenario. This is obtained from watching videos, scary movies, comedic act etc and are recorded without explicitly mentioning that they are being monitored. The following contains all the datasets that have been used to find the best possible results for facial expressions.

1.3 Research Contribution

In a nutshell, the following contributions were made using this study:

- We introduced created a new model for Facial Expression recognition called MEDNet using a customized Dataset which is capable of identifying 3 new expressions "Annoyance", "Irritation" and "Confusion". In order to increase the accuracy of the training, we used a cascaded Local Binary Pattern model which fed the images for training in the Convolution NN. 3.3. We implemented this method to accelerate the face and expression recognition process by looking for contrast in the image.
- In order to train the the huge dataset of over 1,00,000 images, we proposed a new pixel binning algorithm, which cuts down on the compute time by a significant amount, shown in 4.
- A lot of the images in our custom dataset named "VisageEmotioNet" had busy backgrounds. It was necessary that we remove them in order to remove noise from the dataset as far as we can. So we implemented a background separation algorithm to detect faces in the images and crop them according to need
- To correctly identify the facial muscle movements which connote to the Facial Expression, we implemented Mediapipe on the facial images. Our entire research can be applied to static images, video frames or in real-time video feed.

1.4 Thesis Organization

Our thesis here consists of an introductory portion, previous references in this field of research, different methodologies we have applied to get our desired result, and finally discussion and analysis of the result. The introduction contains a proper introduction to our thesis, our motivation behind our research, the novelty of the contributions from our research, and how the thesis is organized for transparency. The remainder of the paper is discussed in the following. Chapter 2 shows the previous research done by other fellow research enthusiasts. It mostly highlights the description of the data-sets and methodologies used to create large image databases. Chapter 3 deep-dives into our own image database VisageEmotioNet, our own model MEDNet for Micro-Facial Expression recognition and how different methods for data-preprocessing, feature extraction, classification, and testing were approached in order get our intended results. Chapter 4 is where we test our claims using various parameters and analysis metrics and compare MEDNet with VisageEmotioNet against popular face models like VGGFace and FaceNet. Chapter 5 draws the closing remarks of the paper.

Chapter 2

Related Work

Facial expression recognition is a vital research area within the broader field of computer vision and affective computing[4]. Early feature-based methods included the use of geometric features such as distances between key facial landmarks. Feature-based techniques have evolved to incorporate both geometric and appearance-based features. For instance, the Local Binary Pattern (LBP) (Ahonen et al., 2006) and Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005) have been employed to capture texture and gradient information, respectively[66].

The application of machine learning algorithms, particularly Support Vector Machines (SVMs), gained popularity in facial expression recognition. Works such as Bartlett et al. (2005) demonstrated the effectiveness of SVMs in classifying facial expressions based on extracted features. These approaches marked a significant shift towards data-driven recognition systems.

To facilitate the development and evaluation of facial expression recognition systems, several facial expression databases have been created. The Cohn-Kanade database (Kanade et al., 2000) and the MMI Facial Expression Database (Pantic et al., 2005) are examples of widely-used datasets that contain labeled facial expressions captured under controlled conditions.

One foundational work in the study of micro-facial expressions is Ekman and Friesen's (1978) development of the Facial Action Coding System (FACS). FACS serves as the basis for categorizing facial muscle movements, enabling a systematic approach to expression analysis. It has been instrumental in training human coders and forms the theoretical foundation for automated micro-expression recognition systems. Micro-expression databases are essential for training and evaluating detection algorithms. Matsumoto and Hwang (2011) contributed significantly to this area by creating databases that contain authentic micro-expressions, enabling researchers to develop and validate recognition methods against real-world data. Hong, and Moilanen (2011) explored the use of dynamic textural patterns for micro-expression recognition[6]. This approach leverages spatial and temporal information to identify subtle facial movements, which are characteristic of micro-expressions. Such techniques are crucial for distinguishing micro-expressions from macro-expressions. Pfister, Li, Zhao, and Pietikäinen (2011) applied machine learning techniques for recognizing spontaneous micro-expressions. Their work demonstrated the potential of algorithms to automatically detect micro-expressions in video sequences, laying the groundwork for automated systems. Jack, Garrod, Yu, Caldara, and Schyns (2012) delved into the cross-cultural aspects of micro-expression recognition. Their

research highlighted the importance of considering cultural differences in micro-expression analysis, shedding light on the variability of expressions across different populations. Li and Deng (2018) addressed the critical need for real-time micro-expression spotting in video data. Their work advanced the practical application of micro-expression recognition, particularly in security and deception detection contexts. Later in 2020, they conducted a comprehensive review of deep learning techniques applied to micro-expression recognition. Their analysis showcased the effectiveness of deep neural networks in capturing intricate patterns and features from micro-expressions, achieving state-of-the-art results. Huang and Wang provided an extensive review of micro-expression detection applications, encompassing fields such as security, psychology, and human-computer interaction in the same year. They also highlighted ethical considerations surrounding privacy and surveillance in the context of micro-expression analysis.

Micro-expression datasets can be categorized into acted or spontaneous examples. Acted micro-expression samples are elicited by the actors themselves after they're told what the expression entails. For spontaneous samples, subjects' emotions are stimulated in real-time or from video. Micro-expression samples elicited spontaneously give a true picture of what it really is when compared with the acted samples. Existing micro-expression database samples that were acted include USFHD database and Polikovsky's database (Polikovsky et al., 2014) while spontaneous ones include Spontaneous Micro-expression database (SMIC), Chinese Academy of Sciences Micro-expression database (CASME) and its updated version, CASME II, AffectNET database, which consists of nearly 1 million high quality images, RAF-DB, making a great approach to negative valence-arousal study etc.

Micro-expression[14] samples were selected based on recordings that had a total duration of less than 500 milliseconds or an onset duration of less than 250 milliseconds. Their labelling was done with the similar criteria that are the same that of ordinary facial expressions. The micro-expression samples consist of seven classes which includes Neutral, Happiness, Sadness, Anger, Disgust, Fear and Surprise. Micro-expression samples from CASMEII were already pre-processed as model. The video sample rate was at 200fps. The variation in the number of samples used for the two set of experiments is as a result of some samples whose coding did not include their correct apex, onset and offset labels. Details on the number of samples for each micro-expression class are presented in Table 1. Pre-processing involves conversion of frames from RGB into grey-scale images.

Early databases[19] of facial expressions such as JAFFE, Cohn-Kanade MMI, and MultiPie were captured in a lab-controlled environment where the subjects would portray different facial expressions. This approach was highly sought after due to the quality of the database being regulated. But it removed the spontaneity aspect of random expressions. Thus, capturing spontaneous expression became a trend in the affective computing community. Examples of these environments are recording the responses of participants' faces while being stimulated. Datasets like DISFA, AM-FED 15, performing laboratory-based emotion inducing tasks in datasets like Belfast [16]). These databases often capture multi-modal affects such as voice, biological signals, etc. and usually dynamic expressions are captured. However, the diversity of these databases is limited due to the number of subjects, financial aspect and environmental conditions.

So a demand to develop systems that are based on natural, unposed facial expres-

sions enveloped the community. Researchers paid attention to databases like in the wild Acted Facial Expressions in the Wild (AFEW) database. AFEW contains 330 subjects aged 1 to 77 years and addresses the issue of temporal facial expressions in the wild. SFEW is created by selecting some frames of AFEW. SFEW covers unconstrained facial expressions, different head poses, age range, occlusions, and close to real world illuminations. However, it contains only 700 images, and there are only 95 subjects in the database.

The Affectiva-MIT Facial Expression Dataset (AM-FED) database [15] contains 242 facial videos (160K frames) of people watching Super Bowl commercials using their webcam. This database was quite automatic in nature and led the research of dynamically created databases according AM-FED is a great resource to learn faces in the wild. However, it was limited as there is not a huge variance and there are only a few subjects in the database.

The FER-Wild [20] database contains 24,000 images that are obtained by searching for emotion-related terms from three search engines. Two human labelers worked on labeling the images into six basic expressions and neutral. Comparing with FER-2013, FER-Wild images were higher in resolution but lacked deficiency in terms of images of certain emotions like Disgust and Fear.

The EmotioNet[33] consists of one million images of facial expressions downloaded from the Internet by selecting all the words derived from the word “feeling” in WordNet. These images were then automatically annotated with AUs by using Kernel Subclass Discriminant Analysis (KSDA). The KSDA-based approach was trained with Gabor features centered on facial landmark with a Radial Basis Function (RBF) kernel. Images were labeled as one of the 23 emotional subclass categories defined in based on AUs. This was unique in nature as it delved into the nature of compound emotions. For example, if an image has been annotated as having AUs 1, 2, 12 and 25, it is labeled as happily surprised. A total of 100,000 images (10 percent of the database) were manually annotated with AUs by experienced coders. The proposed AU detection approach was trained on CK+, DISFA, and CFEE databases, and the accuracy of the automated annotated AUs was reported about 80 percent on the manually annotated set. EmotioNet is a novel resource of FACS model in the wild with a large amount of subject variation. However, it lacks the dimensional model of affect, and the emotion categories are defined based on annotated AUs and not manually labeled.

On the other hand, some researchers developed databases of the dimensional model in the continuous domain. Examples of these databases are Belfast, RECOLA, Affectiva-MIT Facial Expression Dataset (AM-FED), and recently published Aff-Wild Database which is the first database of dimensional model in the wild. Most recent iteration of the model is AffectNET.

The Belfast database contains recordings 5 – 60 seconds length of mild to moderate emotional responses of 60 participants to a series of laboratory-based emotion related tasks. The recordings were labeled by information on self-report, gender and the valence in the continuous domain. The arousal dimension was not a requirement in Belfast database. While the portrayed emotions are natural and spontaneous, the tasks have taken place in a relatively artificial setting of a laboratory where there was a control on lighting conditions, and head poses. These results were relatively binding for these reasons.

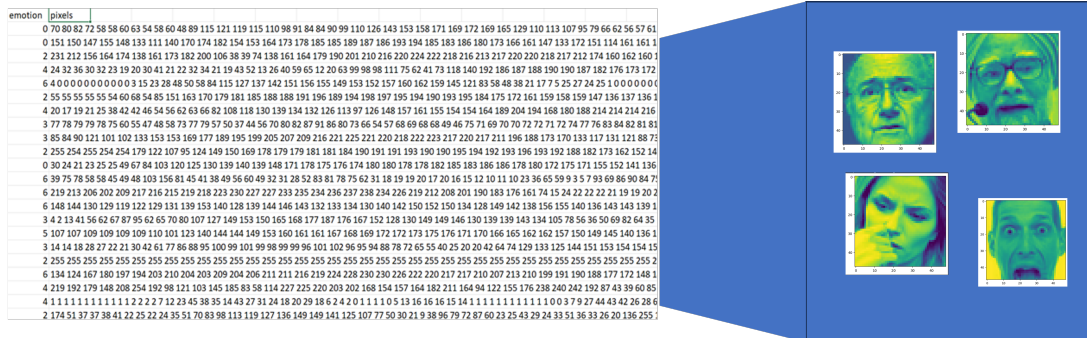
The Database for Emotion Analysis using Physiological Signals (DEAP) consists of

spontaneous reactions of 32 participants in response to one-minute-long music video clip. The EEG, peripheral physiological signals, and frontal face videos of participants were recorded, and the participants rated each video in terms of valence, arousal, like/dislike, dominance, and familiarity. Correlations between the EEG signal frequencies and the participants ratings were investigated, and three different modalities, i.e., EEG signals, peripheral physiological signals, and multimedia features on video clips (such as lighting key, color variance, etc.) were used for binary classification of low/high arousal, valence, and liking. DEAP is a great database to study the relation of biological signals and dimensional affect, however, it has only a few subjects and the videos are captured in lab controlled settings.

The RECOLA benchmark contains videos of 46 participants that participated in a video conference completing a task which needed team collaboration. Different multi-modal data of the first five minutes of interaction, i.e., audio, video, ECG and EDA, were recorded continuously and simultaneously. Six annotators labelled arousal and valence. Self-Assessment was done using a manikin and then questionnaire was carried out before and after the task. RECOLA is an amazing database, however, it contains only 46 subjects and the videos were captured in the lab-controlled settings.

The Aff-Wild Database is one of the largest database for measuring continuous affect in the valence- arousal space “in-the-wild”. More than 500 videos from YouTube were collected. Subjects were performing various day-to-day activities which can attributed to their stimulant. Aff-Wild is a great database of dimensional modeling in the wild that considers the temporal changes of the affect, however, it has a small subject variance, i.e., it only contains 500 subjects.

The Facial Expression Recognition 2013 (FER-2013)[50] database was introduced in the ICML 2013 Challenges in Representation Learning [19]. The database was created using the Google image search API that matched a set of 184 emotion-related keywords to capture the six basic expressions as well as the neutral expression. Images were resized to 48x48 pixels and converted to grayscale. Human annotators labelled the images painstakingly. The resulting database contains 35,887 images most of which are in the wild settings. FER-2013 is currently one of the the biggest publicly available facial expression database in the wild settings, enabling many researchers to train machine learning methods such as Deep Neural Networks (DNNs) where large amounts of data are needed. In FER-2013, the faces are not registered, a small number of images portray disgust (547 images), which causes underfitting, and unfortunately most of facial landmark detectors fail to extract facial landmarks at this resolution and quality. Coding with this database was also extremely tough due to the need for the rendition of the images being converted back into an image afterwards. In addition, only the categorical model of affect is provided with FER-2013.




The emotions are labelled with numbers -
0: 'Anger'
1: 'Disgust',
2: 'Fear'
3: 'Happiness'
4: 'Sadness'
5: 'Surprise'
6: 'Neutral'

Figure 2.1: Fer2013 Dataset

Real-world Affective Faces Database (RAF-DB) is a large-scale facial expression database with around 30K great-diverse facial images downloaded from the Internet. Based on the crowdsourcing annotation, each image has been independently labeled by about 40 annotators. Images in this database are of great variability in subjects' age, gender and ethnicity, head poses, lighting conditions, occlusions, (e.g. glasses, facial hair or self-occlusion), post-processing operations (e.g. various filters and special effects), etc. RAF-DB has large diversities, large quantities, and rich annotations, including 29672 number of real-world images, a 7-dimensional expression distribution vector for each image, two different subsets: single-label subset, including 7 classes of basic emotions; two-tab subset, including 12 classes of compound emotions, 5 accurate landmark locations, 37 automatic landmark locations, bounding box, race, age range and gender attributes annotations per image, baseline classifier outputs for basic emotions and compound emotions.

image	label
train_00001_aligned.jpg	5
train_00002_aligned.jpg	5
train_00003_aligned.jpg	4
train_00004_aligned.jpg	4
train_00005_aligned.jpg	5
train_00006_aligned.jpg	1
train_00007_aligned.jpg	5
train_00008_aligned.jpg	4
train_00009_aligned.jpg	4
train_00010_aligned.jpg	1
train_00011_aligned.jpg	4



Labels –
1: 'Neutral',
2: 'Fear'
3: 'Sadness'
4: 'Happiness'
5: 'Surprise'
6: 'Disgust'

Figure 2.2: RAFDB Dataset

FEER database[52] contains six basic emotions (happiness, surprise, anger, fear, disgust, and sadness) of normalized (average mean reference) data and collected from 85 undergraduate university students (55 male; 30 female) aged between 20 - 27 years with a mean age of 24.5 years. A built-in face time HD camera in Apple Mac Pro with a resolution of 2560×1600 at 227 pixels per inch is used to collect the facial images in a controlled environment (25°C room temperature with 50 Lux lighting intensity) at 30 frames per second. All the subjects are seated comfortably in a chair in front of the camera and the distance between the subject face to the camera is 0.95m. A computerized PowerPoint slides are used to instruct the subjects to express the facial emotional expression by looking into the International Affective Picture System (IAPS) images of six different emotions. The data file contains 11 columns (10 columns for 10 markers and the last column represents the label of emotion) and 190968 rows. In the file (labels), 0 refers to Angry, 1 refers to Disgust, 2 refers to Fear, 3 refers to Sad, 4 refers to Happy, and 5 refer to S to Surprise. Each emotion has 10 trials and each trial has a duration of 6 sec. In between the emotional expressions, 10 sec of break is given to the subjects to feel calm by showing natural scenes.

P_e1	P_e2	P_e3	P_e4	P_m1	P_m2	P_m3	P_m4	Pm6	P_m7	emotion
0.349199	0.283633	0.332759	0.216879	0.428061	0.307376	0.464403	0.477004	0.379559	0.555433	0
0.311972	0.203816	0.327639	0.225092	0.308788	0.248257	0.379399	0.293015	0.404383	0.451443	0
0.310054	0.200216	0.328004	0.224183	0.309334	0.247905	0.377769	0.290631	0.40143	0.449153	0
0.311457	0.203679	0.328945	0.224877	0.311774	0.248453	0.380679	0.290603	0.402368	0.450308	0
0.310568	0.198933	0.320973	0.215203	0.308446	0.247227	0.377342	0.288206	0.401431	0.448102	0
0.311817	0.203123	0.319028	0.215467	0.310794	0.247967	0.378622	0.289461	0.404271	0.449131	0
0.311549	0.198783	0.316427	0.210112	0.306091	0.247268	0.375577	0.285587	0.402586	0.446836	0
0.319652	0.210553	0.338185	0.235335	0.312513	0.248701	0.381375	0.298008	0.404539	0.454884	0
0.318716	0.211579	0.335963	0.233747	0.313535	0.248628	0.381859	0.299646	0.406257	0.455996	0
0.316721	0.207234	0.33329	0.231473	0.310286	0.247663	0.38048	0.298622	0.404977	0.455206	0
0.319019	0.210175	0.336738	0.23403	0.309772	0.24796	0.380736	0.296557	0.404529	0.454491	0
0.319061	0.210417	0.336468	0.234214	0.309545	0.247735	0.380556	0.297235	0.404439	0.454341	0
0.318347	0.209864	0.336248	0.233637	0.310249	0.248086	0.381079	0.297789	0.403858	0.455398	0
0.318253	0.209299	0.335253	0.233058	0.31086	0.24841	0.381704	0.299174	0.40593	0.455021	0
0.317932	0.208658	0.334497	0.232298	0.311582	0.248548	0.381571	0.300192	0.40614	0.455806	0
0.316768	0.205297	0.331124	0.227113	0.308582	0.248987	0.380487	0.29852	0.404997	0.454688	0
0.317025	0.207411	0.334123	0.230965	0.311713	0.249132	0.379463	0.300383	0.407172	0.455425	0
0.316583	0.205096	0.331151	0.22999	0.310549	0.246317	0.380679	0.297557	0.406804	0.452787	0
0.315613	0.207636	0.334222	0.230005	0.311041	0.248382	0.381237	0.300352	0.410105	0.455768	0



Labels are defined as follows–
1: 'Angry',
2: 'Disgust'
3: 'Sad'
4: 'Happy'
5: 'Surprise'
6: 'Disgust'

Figure 2.3: FEER Dataset

The newest addition to the facial expression dataset is the AffectNET Dataset[36] introduced by Mollahosseini et al. in AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild, which arguably has the biggest array of Expression based data of them all, making it one of the main areas of concern for our research study here. The most unique thing about this database is that it searched for terms on the internet that have never been sought after. Emotion-related keywords were combined with words related to gender, age, or ethnicity, to obtain nearly 362 keywords in the English language such as “joyful girl”, “blissful Spanish man”, “furious young lady”, “astonished senior”. These keywords are then translated into five other languages: Spanish, Portuguese, German, Arabic and Farsi.

Therefore, the list of English queries was provided to native non-English speakers who were proficient in English, and they created a list of queries for each emotion in their native language. Due to this huge variation of the data query, the results that came back were not really one to one with the six basic emotions, rather an amalgamation, which made generalization of the data to be quite unique. The criteria for high-quality queries were those that returned a high percentage of human faces showing the intended queried emotions rather than drawings, graphics, or non-human objects. A total of 1250 search queries were compiled and used to crawl the search engines in our database. Search engines such as Google, Yahoo, Baidu and Yandex were considered. A total of 1,800,000 distinct URLs returned for each query were stored in the database. The OpenCV face recognition was used to obtain bounding boxes around each face. A face alignment algorithm via regression local binary features was used to extract 66 facial landmark points. The average image resolution of faces in AffectNet are 425 X 425 with STD of 349 X 349 pixels. Microsoft internal cognitive face API was used to extract these facial attributes on 50,000 randomly selected images from the database. According to MS face API, 49 percent of the faces are men. The average estimated age of the faces is 33.01 years with the standard deviation of 16.96 years. In particular, 10.85, 3.9, 30.19, 26.86, 14.46, and 13.75 percent of the faces are in age ranges [0, 10), [10, 20), [20, 30), [30,

40), [40, 50) and [50, -), respectively. MS face API was able to detect individual facial features. 9.63 percent of the faces wear glasses, 51.07 and 41.4 percent of the faces have eye and lip make-ups, respectively. In terms of head pose, the average estimated pitch, yaw, roll are 0.0,-0.7, and -1.19 degrees, respectively.

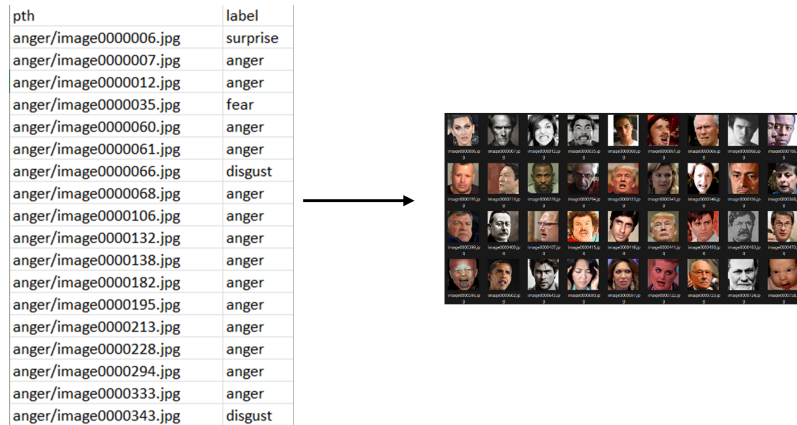


Figure 2.4: AffectNET Dataset

Alongside with our goal of new training approach and optimizations in many areas of our research, we have attempted to add 3 new emotional analysis "Annoyance", "Irritation" and "Confusion". Hence, we created a new Facial Expression Dataset called "VisageEmotioNet", which contains an overall 31503 images with special focus on 500 images that are dedicated to training the 3 stated new expressions. Among these images, 137 images were allocated for performing training computation on "Annoyance", 145 images for "Irritation" and 218 images were dedicated for "Confusion" Emotional Analysis.

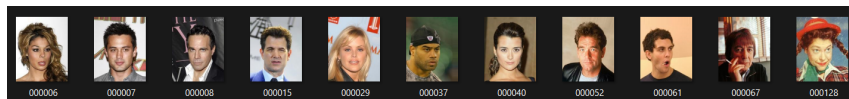


Figure 2.5: VisageEmotioNet Dataset

The sources for these images were through image scraping through various social media and movies. To completely crop optimize the training parameters, we removed the background through Background Separation and consolidated the image size through Pixel Binning.

Macro-Expression Datasets		Micro-Expression Datasets	
Spontaneous Datasets	In-the-wild	Spontaneous Datasets	In-the-wild
EB+, TAVER, RAVDESS, GFT, SEWA, BP4D+ (MMSE), BioVid Emo, 4D CCDb, MAHNOB Mimicry, B3D(AC), CK+, AvID, AVIC, DD, SAL, HUMAINE, OPEN-EmoRec-II, AffectNET, FERG-DB, AVEC'14, BP4D-Spontaneous, DISFA, RECOLA, AVEC'13, CCDb, DynEmo, DEAP, SEMAINE, MAHNOB-HCI, UNBC-McMaster, CAM3D, EmoTABOO, ENTERFACE, UT-Dallas, RU-FACS, MIT, UA-UIUC, AAI, Smile dataset, iSAFE, ISED, Fer2013 Dataset, CASMEII Dataset	FERWild, Vinereactor, CHEAVD, HAPPEI, AM- FED, FER- 2013, AFEW, Belfast induced, SFEW, VAM- faces, RAF-DB, Aff- Wild2, AMFED+, AffectNet, AFEW-VA, Aff-Wild, EmotioNet,	CASME II, CASME, SAMM, CAS(ME)2, Silesian deception, SMIC-E, SMIC, Canal9, YorkDDT, AffectNET	MEVIEW

Figure 2.6: Categorization of Macro and micro Expression Datasets

Amongst all of the datasets mentioned for Macro-expressions Datasets can be used to train Micro-expressions models as well. However, some of them lack the proper optimizations to catch the subtleties of the facial data in a shorter range of time. In summary, the literature in micro-facial expression[38] detection encompasses a wide range of topics, from foundational work on expression coding systems to the latest advances in deep learning-based recognition techniques. This diverse body of research underscores the growing importance and interdisciplinary nature of micro-expression analysis, serving as a strong foundation for the present study.

Chapter 3

Methodologies

3.1 Top Level View of MEDNet

The proposed MEDNet scheme is an aggregation of an effective subtle expression recognition combined with a unique model created using novel feature classification technique, which is categorized into five main phases: (1) Expression based facial classification dataset training, (2) Feature selection using a cascaded Local Binary Pattern-CNN architecture, (3) Data preprocessing using a proposed Background Separation and Pixel Binning Approach, (4) Ensemble of facial expression dataset that have been fed through the previously stated LBP-CNN layer, (5) Model created and implemented in real-world and pre-recorded media using a face identifier for muscle-movement detection.

In order to implement the MEDNet model for training and testing, four datasets have been applied, one of which 2.5 has been assembled by us. All four datasets have been used to in the MEDNet architecture, except the fer2013, which didn't require our proposed pixel binning algorithm to be applied on it. A 3-layer Convolutional Neural Network has been created with a LBP feed-through for the identification of contrast detection at the edges of the faces to accelerate the emotional analysis.

The consequent sections explain the methodological part of the proposed model clearly. In order to improve transparency on the later sections, List of Symbols can be referred to and proper references have been added.

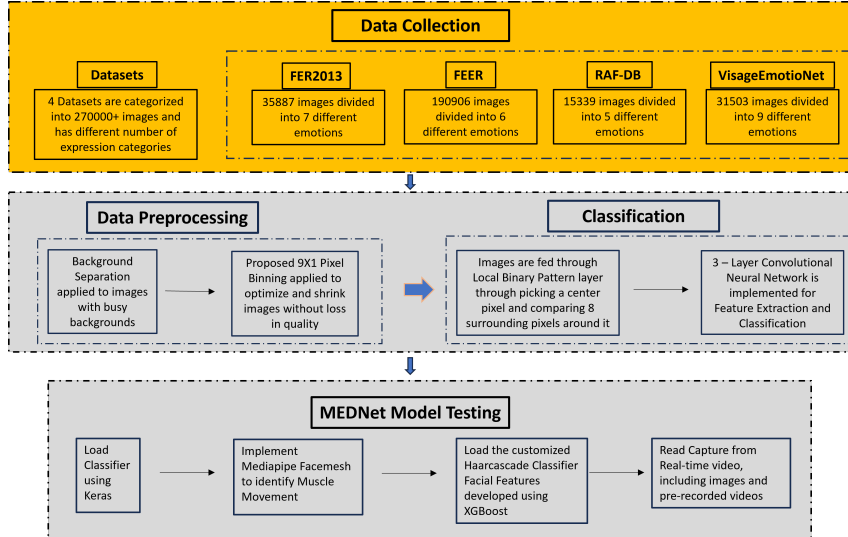


Figure 3.1: Top Level View of Proposed MEDNet Architecture

- Dataset Distribution** shows how many images and datasets we are working with. We are attributing to 4 datasets showing a number of variations in the images. We have our very own VisageEmotionNet that has 9 different emotions distributed over 31503 images, Fer2013 has 35887 images distributed over 7 different emotions, FEER has 190 thousand images over 6 emotions and finally RAF-DB has 15339 images over 5 different emotions. It is explained in complete in chapter 3.2 Data Collection.
- Data Preprocessing** includes a number of steps that are explained in chapter 3.4 - 3.6. The synopsis is explained as:
 - *Background Separation* was utilized to remove busy background during training. All of our collected images were not just "Faces" and needed some background removal techniques to isolate the face
 - We have nearly 270,000 images in total that needed to be trained. Computational Cost aside, this would need a sheer amount of time to parameterize into the training window. Hence we implemented a *pixel binning* algorithm to convert 9 pixels into one pixel which will require significantly less computational time.
 - The *LBP* layer was implemented to help the model find the facial features faster when trained through the *Convolutional Neural Network*.
 - *Haarcascade Classifier* was implemented with the XGBoost classifier for identification of faces in frames of videos or images and real-time camera feed
- MEDNet Model Testing** was tested with our exported model. We loaded our model to identify expressions and Mediapipe was implemented to find Muscle Movement in faces for subtlety in response.

We elaborated on this in the chapters 3.7 to 3.10

3.2 Data Collection

Fer2013[28] was a facial database created by Goodfellow et al to promote research on FER, which was collected from Kaggle. FER2013 dataset consists of 35887 grayscale images with sizes (48x48), all images are the cropped faces, FER2013 dataset has the basic seven expressions, are "angry, disgust, fear, joy, neutral, sad, surprise". The challenges of these images are (have a small size and the low resolution), in addition to being highly imbalanced, because the number of images is different from one emotion to others, making it hard to obtain a model that can work well and accurately for all expressions, also some of the images are rotated, occluded, and with various illumination.

Expression	Data Count
Angry	958
Disgust	111
Fear	1,024
Happy	1,774
Neutral	1,233
Sad	1,247
Surprise	831

Table 3.1: Number of Images per Expression (Fer2013)

The Real-world Affective Faces Database (RAF-DB)[30], created by Li, Shan and Deng, Weihong and Du, JunPingin their very well known paper "Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild", is a dataset that contains 12271 facial images tagged with basic or compound expressions by 40 independent taggers. Images in this database are 100x100 and of great variability in subjects' age, gender and ethnicity, head poses, lighting conditions, occlusions, (e.g. glasses, facial hair or self-occlusion), post-processing operations (e.g. various filters and special effects), etc.

Expression	Data Count
Happy	4772
Sad	1982
Angry	705
Neutral	2524
Surprised	1291
Fear	281
Confused	717

Table 3.2: Number of Images per Expression (RAF-DB)

FEER Dataset has been constituted by 2560×1600 at 227 pixels per inch is used to collect the facial images in a controlled environment (25°C room temperature with 50 Lux lighting intensity) at 30 frames per second. The researchers of this dataset has annotated the dataset using 7 parameters ($P_e[1-4]$) and ($P_m[1-3]$) for feature

extraction. The distinct values for the extraction has then been allocated to an emotion and was applied to classify 190,968 different results.

Parameters	Features Extracted
P_e1	68714
P_e2	65130
P_e3	71651
P_e4	70581
P_m1	73563
P_m2	73794
P_m3	61376

Table 3.3: Number of Images per Expression (FEER-DB)

VisageEmotioNet dataset is quite varying in nature amongst all the dataset, containing a number of images of various races and ages in various lighting and environment conditions.

In fact, it contains a magnitude of Dimensional data for exploration solely dedicated to the research of Cognitive-facial-expression. Crowd-sourcing services like Amazon Mechanical Turk are fast, cheap and easy approaches for labeling large databases. However, they do not ensure the quality desired and have been avoided with an alternative being the employment of 12 full-time and part-time annotators at the University of Denver to label the database. A total of 31500 images were collected which was divided into 9 categories.

A comprehensive study including the definition of the categorical and dimensional models of affect with some examples of each category, valence and arousal was done. Three training sessions were provided in which and necessary feedback was given on both the categorical and dimensional labels. In addition, we tagged the images that have any occlusion on the face. If anyone in the images wore glasses, but the eyes were visible without any shadow, it was not considered as occlusion.

Eleven discrete categories were defined in the categorical model of VisageEmotioNet as: Annoyance, Surprise, Satisfaction, Neutrality, Concentration, Irritation, Confusion, Disgust, Sadness.

Expression	Number
Neutral	5132
Happy	5045
Sad	3430
Surprise	4296
Concentration	789
Disgust	2660
Annoyed	137
Confusion	218
Irritation	145

Table 3.4: Number of Images per Expression (VisageEmotioNet)

3.3 Overall System Model with Description

The proposed approach consists of two main stages: feature extraction and classification. In the feature extraction stage, using background subtractions, we separate the facial features from the image, then we compute LBP descriptors from the input images, capturing local texture patterns at multiple scales. These LBP descriptors are then used as an additional input channel, combined with the standard RGB channels, to form an enriched input representation for the CNN. This should accelerate the image processing pipeline in order to get faster samples. The Background Subtraction methodology is used here because alongside with the dataset, we have trained three new emotions namely, "Annoyance", "Irritation" and "Confusion".

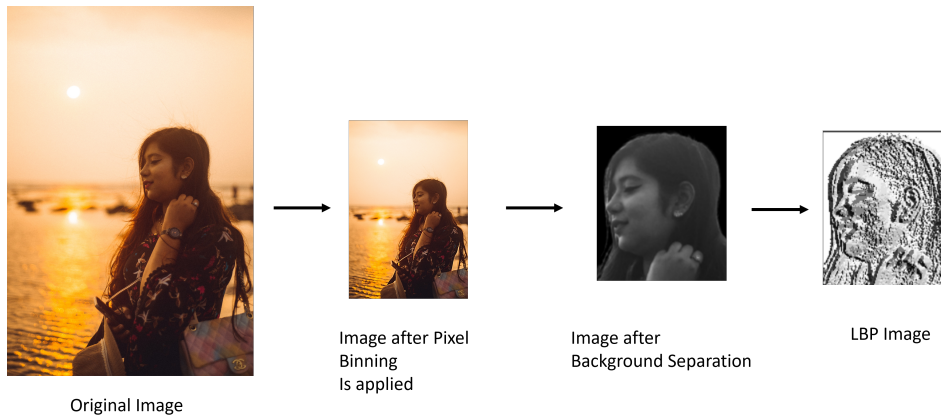


Figure 3.2: Image processing pipeline through Background Separation and LBP

3.4 Dataset Training Approach

To evaluate the proposed approach, extensive experiments are conducted on benchmark datasets, including but not limited to VisageEmotioNet, Fer2013, CelebA and RAF-DB and a custom dataset with various texture classes. We compare the performance of the hybrid CNN-LBP model against traditional CNNs[16], standalone LBPs, and other texture-based classification methods. The evaluation metrics include classification accuracy, precision, recall, and F1-score, providing a comprehensive analysis of the model's performance.

The CNN architecture is carefully designed to accommodate the multi-channel input, allowing it to learn spatial hierarchies of features from both raw pixel data and LBP descriptors. We experiment with different CNN architectures, such as VGG, FaceNet, and Inception, to determine the best combination of network depth and complexity for texture-based image classification. It is constructed as follows:

- **Input Layer:** The first layer of the CNN that receives the input data, which is typically an image or a feature map from a previous layer. The input layer is fed with images from the dataset with the existing datasets alongside with the new images that have been processed using a Background Subtraction and Pixel Binning algorithm.
- **Convolutional Layers:** These layers consist of multiple filters (also called kernels) that slide over the input data to extract features. Each filter is a small

matrix that performs a convolution operation by element-wise multiplication and summation with the input. This operation helps the model learn local patterns and features present in the data. Well there are 4 layers of convolution that have been used for training here, each layer will be 64, 128, 256 and 512 in size. The first 64 convolutional layer will have a Batch Normalization function to maintain proper scaling of the images and the final layer of the CNN will also flatten the 2D arrays into a linear vector. Total trainable parameters are 2,069,959.

- **Activation Function:** After the convolution operation, ReLU activation function was applied element-wise to introduce non-linearity, allowing the CNN to learn complex patterns.
- **Pooling Layers:** Pooling layers reduce the spatial dimensions of the feature maps, reducing computational complexity and extracting more important features. The most common pooling operation is max pooling, which takes the maximum value within a small region.
- **Fully Connected Layers:** After several convolutional and pooling layers, the feature maps are flattened into a one-dimensional vector. This vector is then passed through one or more fully connected layers, similar to a traditional neural network. These layers combine features from different spatial locations and make the model capable of classification.
- **Output Layer:** The final layer of the CNN, which produces the output based on the facial feature performed. For image classification, this layer will have a softmax activation function. It will comprehensively define the expression that is shown either through a video feed, image or a camera.
- **Loss Function:** The loss function quantifies the difference between the predicted output and the ground truth. In image classification tasks, the commonly used loss function is categorical cross-entropy.
- **Optimization:** Our CNN uses one of the variants of Stochastic Gradient Descent called Adam Optimizer to update the model's parameters and minimize the loss function during training.
- **Training:** The CNN is trained using a labeled dataset (input data and corresponding labels) to learn the appropriate weights and biases for each layer. This is accomplished through backpropagation, where the gradients of the loss with respect to the model's parameters are computed and used to update the parameters during optimization.

We have applied our model on students mostly. The labeling was based on emotions that better reflect the emotions shown by the students during studying, alongside with studying the emotions on candid actors or in a simulated environment. The emotion parameters under consideration are "Surprise", "Annoyance", "Satisfaction", "Neutral", "Concentration", "Sadness", "Confusion", "Irritation", "Confusion".

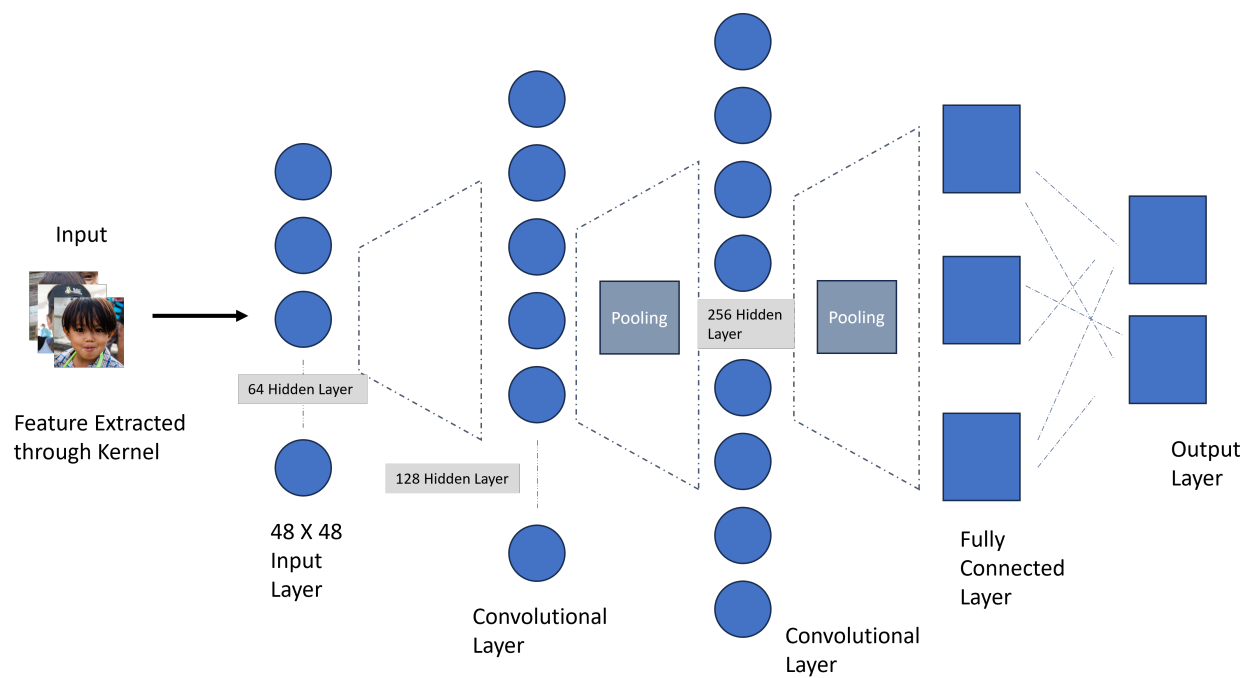


Figure 3.3: CNN

3.5 Data Preparation

Since we are dealing with massive Datasets, we have implemented a few approaches to training:

- The number of images for example, the VisageEmotioNet dataset, is far too large, particularly 450,000 images. So we are training around 31,503 images of VisioEmotioNet. To further reduce the computational cost, a 9x1 pixel merging lossless down-scaling technique that will be named "Pixel Binning" will be applied to further reduce training time. This will be discussed further later on. The similar technique will be applied on the FEER and RAF-DB dataset.
- Fer2013 dataset will not go through any downscaling due to the pictures themselves being smaller in size.
- A custom dataset was created in order to identify additional expressions such as "Annoyance", "Irritation" and "Confusion". These required an extra 500 images that had were scraped from content available online and personal media and was added with the regular expressions. A background separation technique was implemented in order to separate the background from the foreground and pixel binning was applied to increase the computational efficiency.

3.6 Local Binary Pattern

Local binary pattern[8], like pixel binning, applies a very similar equation in terms of analyzing pixels. It picks a central pixel and analyzes the surrounding 8 pixels and assigns a value to it based on the following parameters[13]. We shall use (X,Y)

as the center pixel to explain the position of the pixels in the localized array and a variable called *Pattern* to change the pixel value of the LBP pixel. *Pattern* is initialized to 0. The rules are as follows:

- If it is the center pixel (X, Y) , change the pattern to 1
- If it is located at $(X-1, Y)$, change the pattern after right shifting it to 1 or assigning the value 2 to it
- If it is $(X-1, Y+1)$, change the pattern after right shifting it to 2 or assigning the value 4 to it
- If it is $(X, Y+1)$, change the pattern after right shifting it to 3 or assigning the value 8 to it
- If it is $(X+1, Y+1)$, change the pattern after right shifting it to 4 or assigning the value 16 to it
- If it is $(X-1, Y)$, change the pattern after right shifting it to 5 or assigning the value 32 to it
- If it is $(X+1, Y-1)$, change the pattern after right shifting it to 6 or assigning the value 64 to it
- If it is $(X, Y-1)$, change the pattern after right shifting it to 7 or assigning the value 128 to it

It goes as follows:

Algorithm 1 get lbp pixel (gray image, center, x, y)

```
if neighbouringpixel( $X, Y$ )  $\geq$  CenterPixel then
    neighbouringpixel( $X, Y$ )  $\leftarrow$  1
end if
if neighbouringpixel( $X - 1, Y$ )  $\geq$  CenterPixel then
    neighbouringpixel( $X - 1, Y$ )  $\leftarrow$  (1 * right - shift)1
end if
if neighbouringpixel( $X - 1, Y + 1$ )  $\geq$  CenterPixel then
    neighbouringpixel( $X - 1, Y + 1$ )  $\leftarrow$  (2 * right - shift)1
end if
if neighbouringpixel( $X, Y + 1$ )  $\geq$  CenterPixel then
    neighbouringpixel( $X, Y + 1$ )  $\leftarrow$  (3 * right - shift)1
end if
if neighbouringpixel( $X + 1, Y + 1$ )  $\geq$  CenterPixel then
    neighbouringpixel( $X + 1, Y + 1$ )  $\leftarrow$  (4 * right - shift)1
end if
if neighbouringpixel( $X - 1, Y$ )  $\geq$  CenterPixel then
    neighbouringpixel( $X - 1, Y$ )  $\leftarrow$  (5 * right - shift)1
end if
if neighbouringpixel( $X + 1, Y - 1$ )  $\geq$  CenterPixel then
    neighbouringpixel( $X + 1, Y - 1$ )  $\leftarrow$  (6 * right - shift)1
end if
if neighbouringpixel( $X, Y - 1$ )  $\geq$  CenterPixel then
    neighbouringpixel( $X, Y - 1$ )  $\leftarrow$  (7 * right - shift)1
end if
```

Algorithm 2 lbp (Original Image)

```
convert provided image to gray image
for  $x$  in height of gray image do
    for  $y$  in width of gray image do
        get lbp pixel (gray image, center, x, y)
    end for
end for
```

3.7 Proposed Pixel Binning Approach

Pixel binning has been a growing field of research[2]. Pixel binning typically refers to merging a number of pixels into one pixel. Current implementations include a variety of complex usecases in Smartphone camera technologies. Traditional implementation is the combination of pixel transformation which is to take a number of neighboring pixels and merge them together in a 2x2 or 3x3 pixel binning. But that method can not provide perfect image quality. Our proposed method is capable of extracting information of the pixels using a localized pixel array of 9 pixels and converting them to 1 pixels, retaining the yCbCr or Red-Green-Blue[42] value of the original 9 neighbouring pixels. The color grading of the pixels of the new image will

very closely resemble the original image, only down scaled. The original features like contrast and sharpness shall not be affected.

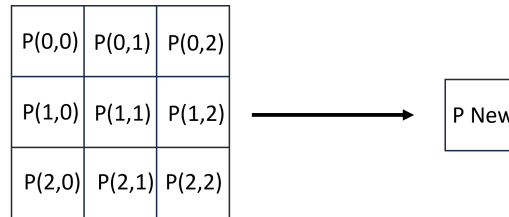


Figure 3.4: 9x1 Pixel Binning

Pixel binning compresses the original image to a 9:1 ratio. To exemplify, it cross checks a number of old pixel and converts them all in a local array. Let's say, Pixel value from P(0,0) to P(2,2) must be downscaled. So we will only create a localized array that will only cover the 9 neighbouring pixels of Pixel P(1,1). P(1,1) will become the P(New) pivot pixel after it had achieve the RGB values of the neighbouring pixels and itself. The local array moves to the neighbouring 3X3 image array and repeats the process until the whole pixel array is binned[10]. As said before, pixel binning is always lossless and the effect of the modelling of the data itself is miniscule. To verify the effect of Pixel Binning, we evaluated the performance before and after pixel binning. The results showed virtually no difference in training.

Algorithm 3 Pixel Binning

$NewHeight \leftarrow \frac{1}{3} * OriginalHeight$
 $NewWidth \leftarrow \frac{1}{3} * OriginalWidth$
for $x, y \leftarrow NewHeight, NewWidth$ **do**
 get RGB value of the pixel(x,y) of the original image
 make average RGB value of the 3x3 original image array
 $NewPixel \leftarrow AverageRGBvalue$
end for

Dataset	Accuracy Before PB (in percentage)	Accuracy After PB (in percentage)	Training Time Ratio of Original image to PB
VisageEmotioNet	98.24	98.16	1:5.06
RAFDB	96.77	95.96	1:6.6
Fer2013	96.12	N/A	N/A
FEER	91.98	92.06	1:4.8

Figure 3.5: Dataset accuracy before and after Proposed Pixel Binning Approach

In order to observe the differences before and after the downscaling of pixel binning, conducted a number of tests. Figure 4.4 shows the advantage this technique has shown us. We have derived an explanation of this figure accordingly:

- We created two models for each datasets, the first model was created using 300 images of every type of emotions before PB(Pixel Binning) is applied and the second model was created using 300 images of the same respective datasets after PB is applied.
- Although after PB, the average degradation of image features were 15.4 percent at max with a 9.7 percent average deviation, the accuracy of training were unaffected, with slight differences that can be attributed to margin of error.
- The main difference is the training time ratio. For VisageEmotioNet, the training time was 17 minutes 31 seconds before PB, which reduced to 3 minutes 37 seconds on the Post-PB datasets. This was generally the trend for all the dataset as the overall reduction in training time was around 5.5 times. Attributing to this, in order to train a massive dataset like VisageEmotioNet, which had 31503 datasets alongside with our 500 custom images would have taken around 30 hours. But we were able to achieve training completion at 6 hours 4 minutes.

3.8 Background Separation

While working with raw iamges we came across a number of concerns. A lot of the images attributed to the following problems:

- Complex Backgrounds: Images with cluttered or complex backgrounds can confound traditional methods, necessitating more advanced techniques.
- Variability: Variations in lighting, facial expressions, and poses pose challenges for consistent background separation.
- Real-Time Processing: In real-time applications, such as video analysis, background separation must be performed swiftly, imposing constraints on computational complexity.

We had to implement a new background separation[32] technique for the stated reasons above. Our proposed Background Separation technique works with clipping the image through evaluation of the clipping plane. The forward plane would be clipped from the background plane. We have used our haarcascade classifier 3.11 here which would remove everything except the face, using the K-nearest Neighbour to achieve this.

KNN algorithm stores all available cases and then it classifies the new samples based on the similarity measure. It is a widely used non parametric technique. In this algorithm, all the training samples are needed to be stored before the classification process. This could be a drawback if a very large data set is used.

N- dimensional numeric attributes are used as the training samples. Each sample represents a point in an n- dimensional space. When any unknown sample needs to be classified, k nearest neighbour classifier searches the pattern space for the k training samples that are closest to the unknown sample. Euclidean distance is used to measure the closeness.

$$d(A, B) = \sqrt{\sum_{n=1}^{\infty} (a_i - b_i)^2} \quad (3.1)$$

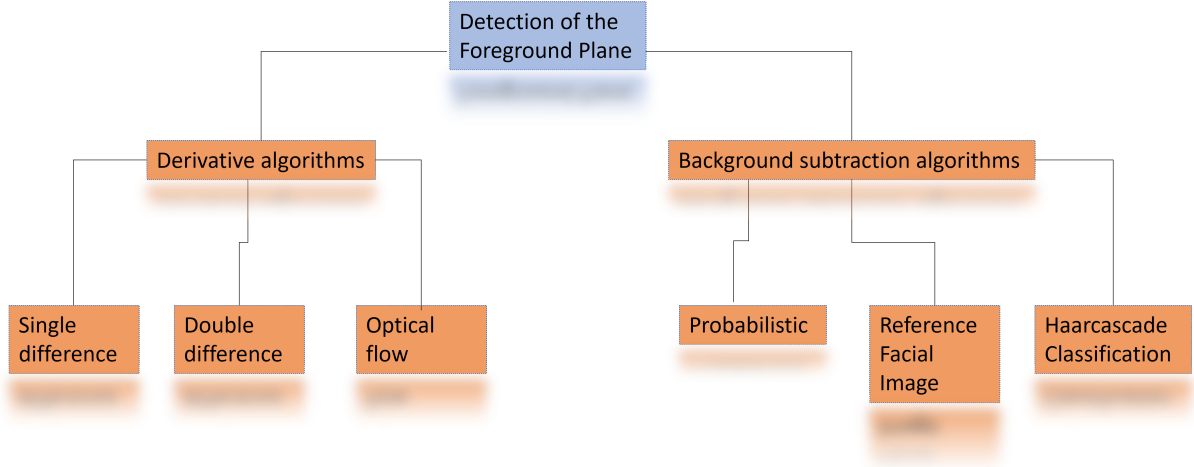


Figure 3.6: Background Separation Flowchart

Background Subtraction will help us get a better representation of the faces in a frame. The image generated will isolate the faces from the dataset. We have applied Background Separation in the following way:

Algorithm 4 Pixel Binning

while CAP is being read **do** *▷ Cap can be a video feed, image or real time camera capture*
 Identify and extract the face from the background
 kernel value of 5x5 is applied to face-image
 resize CAP to 640X480 Resolution *▷ This resolution decreases computational load*
 Apply Erosion and Dilation on the Current Frame
end while

- **Erosion** simply erodes away the boundaries of foreground object, keeping the foreground white. The kernel slides through the image (as in 2D convolution). A pixel valuing 1 or 0 will turn to 1 through this 5X5 kernel convolution and will turn white. Using a 5X5 kernel means that the foreground under the kernel is automatically multiplied by a 5X5 sub-image array of itself. Morphological erosion removes thin floating lines and pixel to make substantive objects stand out. The lines that are not eroded away becomes smaller and thinner as result. Let X be our reference image and S be the structuring element of size 3X3. The Erosion operation is defined by the following equation,

$$X \ominus S = (z | [(\hat{S})_z \cap X]) \in X \quad (3.2)$$

Erosion is a thinning operator, since whatever part is connected to the image is neglected and turned to white using the structuring element S .

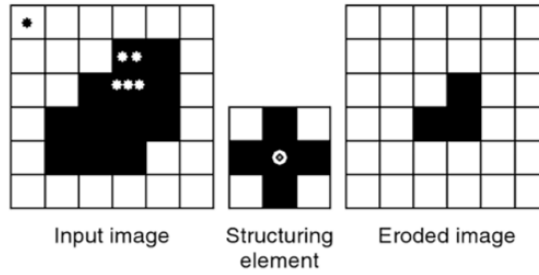


Figure b. Erosion

Figure 3.7: Erosion

- Dilation** does the opposite and simply removes all subpixels under the kernel valuing anything other than 1. This step is required because if we apply erosion, we remove white noise alongside with the shrinkage of our object. Hence we need dilation in order to introduce new pixels that have been eroded away by the Erosion process. Morphological dilation makes way objects more visible and fills in small holes in objects. Lines and objects appear thicker. Let X be our reference image and S be the structuring element of size 3×3 . The dilation operation is defined by equation,

$$X \oplus S = (z | [(\hat{S})_z \cap X]) \in X \quad (3.3)$$

where S is the image rotated about the origin. This equation states that when the image X is dilated by the structuring element S , the outcome element z would be that there will be at least one element in S that intersects with an element in X . So whatever part of the image that coincides with S is expanded.

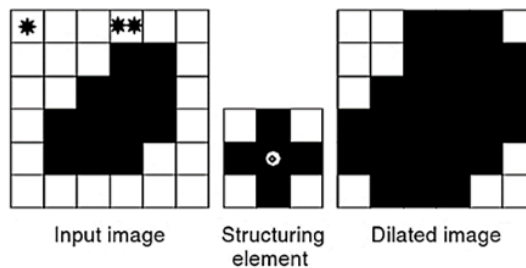


Figure a. Dilation

Figure 3.8: Dilation

The dilation is an expansion operator that magnifies binary objects. Dilation is used majorly for bridging gaps in an image, due to the fact that S is an expanding element for the features of X .

3.9 Mediapipe Facial Movement Detection

In order to correctly identify the facial features[41] that are currently in the frame of reference, we have used Mediapipe[55], A facial muscle movement recognition using Googles implementation of the Facenet.

FaceNet is a start-of-art face recognition, verification and clustering neural network. This 22-layers deep neural network that directly trains its output to be a 128-dimensional embedding. The loss function used at the last layer is called triplet loss.

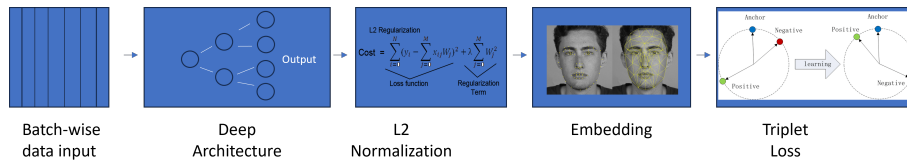


Figure 3.9: Mediapipe Implementation

MediaPipe also offers face tracking capabilities, which can be used to track the same face across consecutive frames in a video stream. This helps maintain consistency and provides the ability to track facial features over time.

There is a total 480 feature tracking dots that are masked onto the reference frame. They are there to detect any sort of muscle movement. This facemesh correctly identifies the lips, eyes, nose and so on. This allows users to better understand and explain why an emotion is being shown and what sort of muscle movements are applied to what sort of emotions. These dots track these expressions after calculating the distance between the dots themselves.

We parameterize tracking confidences to 0.1 as to keep the 3D facemesh on with the highest tracking confidence and keep on trailing the face as long as they can.

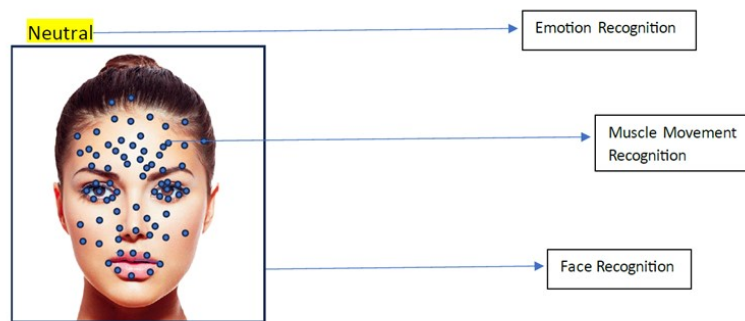


Figure 3.10: MEDnet Classification using Mediapipe

Algorithm 5 Mediapipe Algorithm

min – tracking – confidence ← 0.1

max – tracking – confidence ← 0.1

static – image – mode = *false*

while Capture is being read and face is detected **do**

 Put maxium 480 max facial landmark contours using facemesh

end while

3.10 Haarcascade Classifier implementation

The Haar Cascade classifier[9] is a machine learning-based object detection algorithm that was proposed by Viola and Jones in their 2001 paper "Rapid Object Detection using a Boosted Cascade of Simple Features." It is widely used for real-time object detection, especially for detecting faces in images and videos.

The Haar Cascade classifier works by using a set of simple Haar-like features to represent the patterns of objects. These Haar-like features are rectangular filters that are applied to different regions of an image. Each feature represents a specific pattern, such as edges, lines, or corners.

Haar-like Features: Haar-like features are rectangular filters that are applied at different positions and scales on an image. Each feature represents a specific pattern, such as a horizontal or vertical edge or a change in intensity. These features are used to capture the presence of various patterns in an image.

Integral Image: To efficiently compute Haar-like features, the integral image technique is used. The integral image is a 2D array where each element contains the sum of pixel intensities in the original image up to that point. This technique allows for quick calculation of the sum of pixel intensities in any rectangular region of the image.

The Haar Cascade classifier is trained using a machine learning algorithm. Here we applied the XGBoost classifier to extract the Facial Features.

Cascade of Classifiers: The trained Haar Cascade classifier consists of multiple stages, each containing a set of weak classifiers. Each stage aims to reduce the false positives while maintaining a high detection rate. Weak classifiers are combined using a weighted majority voting mechanism, and the cascade structure allows for efficient rejection of non-object regions in the image. It typically works with Darker and Lighter pixels. Ideally pixels are supposed to be Dark and White, however, values can be assigned to identify the different shades the Dark-White spectrum.

Object Detection: During the object detection phase, the Haar Cascade classifier is applied to the input image using a sliding window approach. The classifier slides the detection window across the image at different scales and positions, applying the Haar-like features to each window. If the combination of weak classifiers in a stage passes a certain threshold, the region is classified as a potential object.

Non-Maximum Suppression: To remove duplicate detections and select the most relevant bounding boxes, non-maximum suppression is applied. This step ensures that only the most confident and non-overlapping detections are retained.

Creating Positive and Negative Sample Description Files:

- Create a positive sample description file listing the filenames of positive images along with the bounding box coordinates.
- Create a negative sample description file listing the filenames of negative images.

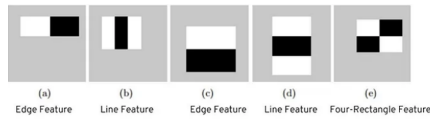


Figure 3.11: Haarcascade Classifier

Out training algorithm to train the positive and negative images were done through annotations of images and then training with XGBoost classifier.

Algorithm 6 Mediapipe Algorithm

- Annotate the positive image and negative images
 - Train using cascaded facial parameter with proper annotations
 - Split the Training and Testing images by 80 to 20
 - Load the XGB classifier and start training
-

Chapter 4

Analysis of Result

4.1 Performance Evaluation Metric

The results were segmented in a number of ways. Firstly, we apply our hybrid LBP-CNN model on the datasets. Performance parameters under considerations are Accuracy of Training Each dataset, Highest Correlation Co-efficient, RMSE Calculation, Recall, Precision and F1-Score Calculation of our LBP-CNN.

The application of LBP-CNN in MEDnet expression has shown very good performance results.. The testing of the Micro-facial expression has been done in three ways – 1. Using Live Feed 2. Using Facial Expression example videos. 3. Static Images.

RMSE Calculation, Recall, Precision and F1-Score Calculation for LBP-CNN was done using the following standardized equations as stated below. These parameters are shown in the tabular form and does not connote to the speed of the recognition, rather the calculation done using Confusion Matrix and the accuracy of prediction. The speed of the recognition will be discussed in a latter section.

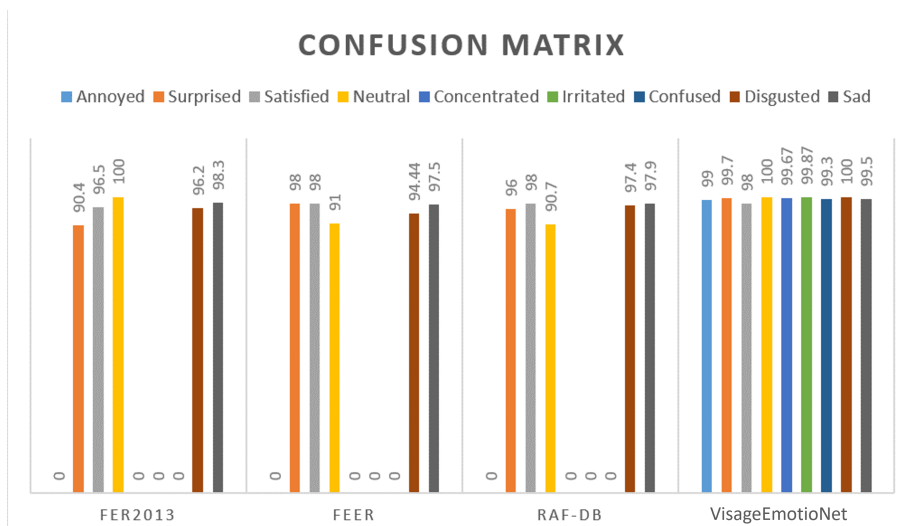


Figure 4.1: Confusion Matrix comparison between various datasets

The confusion matrix have been devised here to understand the correlation of data between the datasets themselves. We labelled our expressions differently from traditional datasets as shown in 4.1

Normalized Confusion Matrix

	Annoyance	Surprised	Satisfied	Neutral	Concentrated	Irritated	Confused	Disgusted	Sad
Annoyance	0.975	0.002	0.001	0.000	0.002	0.005	0.000	0.001	0.002
Surprised	0.003	0.984	0.004	0.000	0.001	0.001	0.001	0.004	0.002
Satisfied	0.006	0.005	0.981	0.000	0.005	0.002	0.000	0.003	0.006
Neutral	0.000	0.001	0.000	0.990	0.011	0.003	0.001	0.002	0.005
Concentrated	0.000	0.000	0.003	0.001	0.978	0.004	0.000	0.000	0.002
Irritated	0.090	0.000	0.002	0.000	0.002	0.980	0.002	0.009	0.002
Confused	0.030	0.003	0.004	0.000	0.000	0.003	0.994	0.007	0.001
Disgusted	0.003	0.003	0.002	0.000	0.000	0.002	0.000	0.973	0.001
Sad	0.001	0.002	0.003	0.000	0.001	0.002	0.001	0.001	0.979

Figure 4.2: Normalized Confusion Matrix of MEDNet Model with VisageEmotioNet Dataset

Fig4.2 shows the normalized confusion matrix of VisageEmotioNet Dataset that was used to create MEDNet model.

Overall, VisageEmotioNet + MEDNet showed a 98.16 percent accuracy across all the classes combined after Pixel Binning was applied .

$$RMSE = \sqrt{\frac{1}{N} \sum_1^N (y_j - \hat{y}_j)^2} \quad (4.1)$$

$$Precision = \frac{TP}{TP + FP} \quad (4.2)$$

$$Recall = \frac{TP}{TP + FN} \quad (4.3)$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4.4)$$

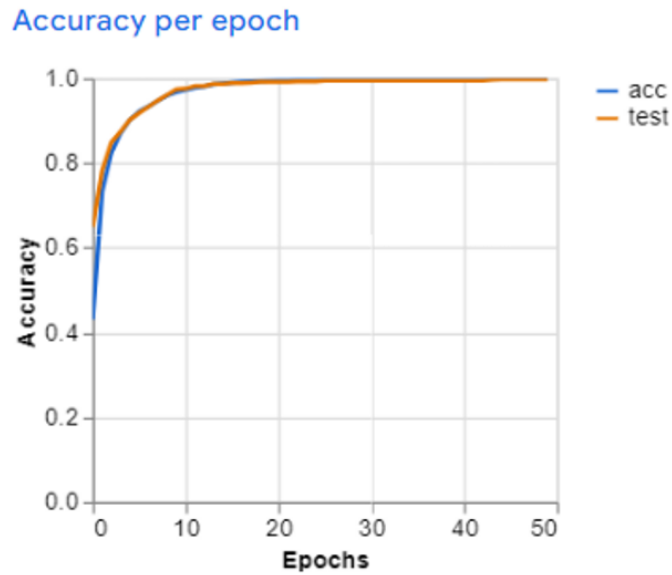


Figure 4.3: Accuracy per epoch

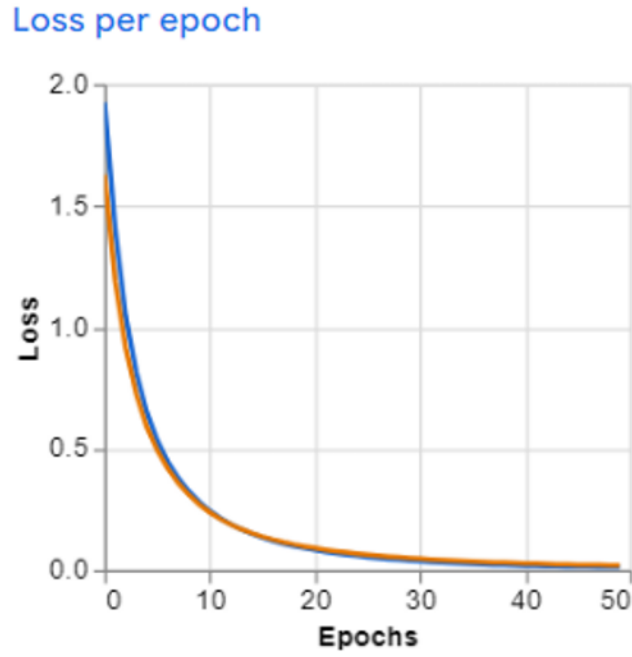


Figure 4.4: Loss per epoch

Accuracy, Precision, Recall and F1 Score derived from the Confusion Matrix have been added. From fig , VisageEmotioNet performs the best with a 98.16 percent accuracy with an RMSE is 0.081.

Expressions	Accuracy	Precision	Recall	F1Score
Annoyed	99.29%	1.0	0.94	0.97
Surprised	98.22%	0.91	0.95	0.93
Satisfied	98.34%	0.93	0.93	0.93
Neutral	98.93%	0.95	0.96	0.96
Concentrated	98.69%	0.95	0.94	0.95
Irritated	98.45%	0.93	0.94	0.94
Confused	98.45%	0.93	0.94	0.94
Disgusted	98.45%	0.93	0.94	0.94
Sad	98.22%	0.93	0.93	0.93

Figure 4.5: VisageEmotioNet Performance Parameters

Expressions	Accuracy	Precision	Recall	F1Score
Surprised	94.93%	0.83	0.93	0.88
Satisfied	96.44%	0.91	0.92	0.91
Neutral	96.06%	0.91	0.88	0.90
Disgusted	97%	0.97	0.88	0.92
Sad	94.93%	0.88	0.87	0.88

Figure 4.6: FEER Performance Parameters

Expressions	Accuracy	Precision	Recall	F1 Score
Surprised	97.64%	0.93	0.95	0.94
Satisfied	97.64%	0.94	0.94	0.94
Neutral	98.23%	1.0	0.92	0.96
Disgusted	98.04%	0.97	0.93	0.95
Sad	97.05%	0.88	0.98	0.93

Figure 4.7: Fer2013 Performance Parameters

Expressions	Accuracy	Precision	Recall	F1 Score
Surprised	92.39%	0.81	0.82	0.82
Satisfied	93.1%	0.80	0.87	0.83
Neutral	94.87%	0.84	0.88	0.86
Disgusted	95.22%	0.95	0.82	0.88
Sad	93.63%	0.84	0.85	0.84

Figure 4.8: RAF-DB Performance Parameters

4.2 Comparative Analysis with VGGface and FaceNet

In order to understand how MEDNet is evaluated with existing popular models like VGGFace and FaceNet[53], we have run some performance evaluation parameters. Our focus is to compare our created model with two of the best performing facial recognition models known as FaceNet and VGGface. The testing methodology is segmented in a few parts, all testing was done using the VisioEmotioNet Dataset

4.2.1 Expression Recognition

Since our primary goal is expression recognition, we have compared all three models for VisioEmotioNet Dataset. This test was conducted on a 1000 image dataset from our VisioEmotioNet dataset. The distribution is 150 images from Annoyed emotions, 90 Surprised, 70 Satisfied, 130 Neutral, 80 Concentrated, 90 Irritated, 80 Confused, 70 Disgusted and 150 Happy images. We observed the following results:

- FaceNet was able to correctly identify 94.01 percent (avg) of the expressions
- VGGFace was able to correctly identify 94.78 percent (avg) of the expressions
- MEDNet was able to correctly identify 96.5 percent (avg) of the expressions

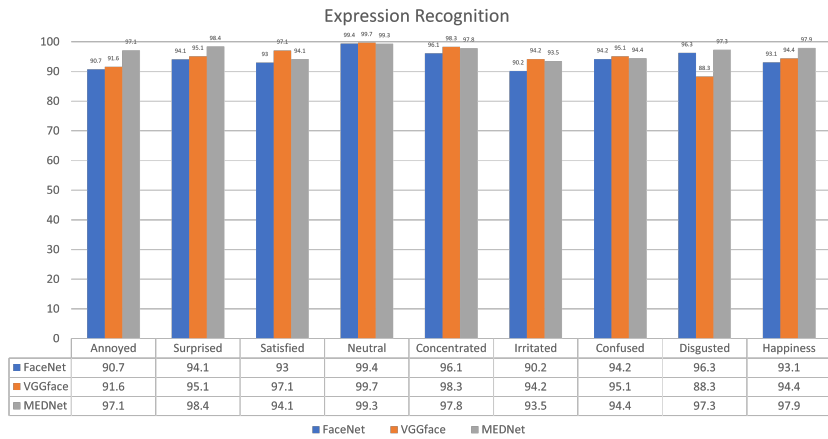


Figure 4.9: Expression Recognition Comparison

4.2.2 Facial Non-Appearance Recognition

Like the previous test, we have selected *1000 images*, except that **380 images** had no faces in it. These images comprised of pictures of landscapes, nature, empty streets etc. The outputs of these tests were parameterized based on two aspects:

- Ability to identify the absence of faces in the frame. Which means testing the models compatibility with finding the non-existence of faces in the **380 images**.
- Co-efficient of ratio analysis of incorrect and correct facial appearance i.e how close to the correctness the **620:380 images** ratio the models can get.

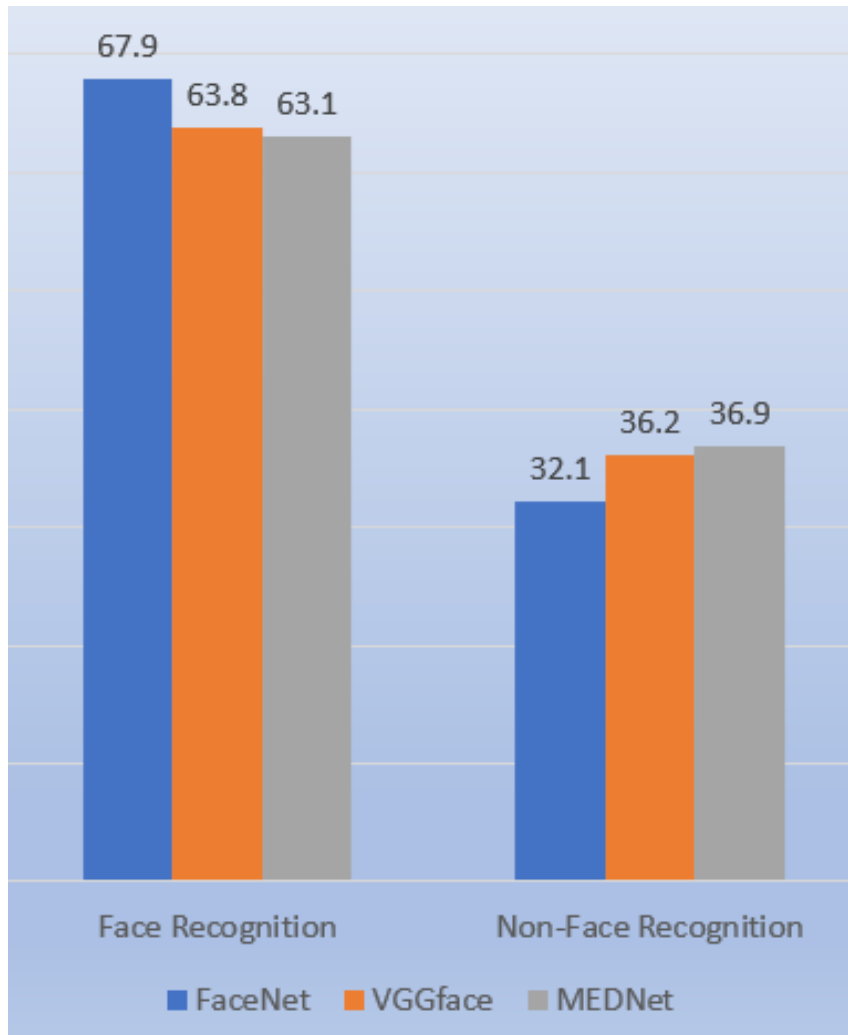


Figure 4.10: Face and Non-Face Recognition for models

Our observations for this test are as follows(fig):

- FaceNet identified **679** images to have faces in it and **321** images without any face, giving it a correction ratio of **2.11**, whereas ideally we were seeking for a ratio of **1.63** i.e **620:380**.
- VGGFace identified **638 images** and **362 images** without any faces were identified by it, which gives us a ratio of **1.76**, a lot closer to **1.63** that we were seeking after.
- MEDNet detected **631 images** and **369 images** for face and non-face embeddings, making a ratio of **1.71**, just **0.8 points** away from our ideal ratio of **1.63**.

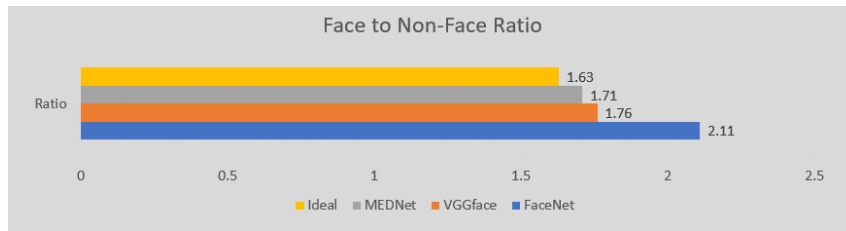


Figure 4.11: Face to Non-Face Ratio

It is to say that these ratios are inline with the research conducted by Yalavarthi Bharat Chandra, Gouru Karthikeya Reddy on their paper ”A Comparative Analysis Of Face Recognition Models On Masked Faces”.

4.2.3 Face and Name Recognition

This test will work as the basis of a face being present on the frame and how precisely the models can recognize the presence and names of the face it has on the frame. The parameters and results for the test are given as follows:

- We created a video file of *1000 static images* with face on them and ran them through the model. We normalized the results within range **95 to 100 percent** since all data points were within this range
- FaceNet was able to recognize **988 images** that had faces on it, and was able to name **992 images**
- VGGface was able to recognize all **1000 images** with faces on it and identified the names of **962 images**
- MEDNet recognized **992** images of facial data and name **991** of them

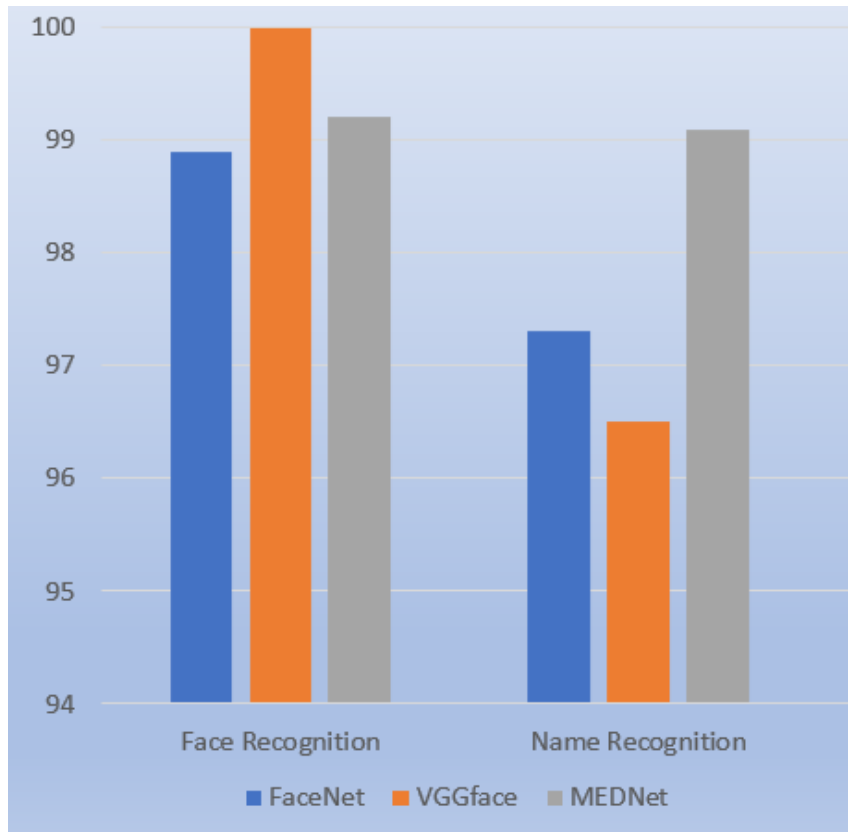


Figure 4.12: Comparison of Models in terms Facial and Nomenclature Data

It is too mention that, MEDnet trained using other Datasets like Fer2013, Feer and RAD-DB were not nearly as accurate and performed recognized 778, 894 and 911 faces in the frames respectively with a 738, 789 and 879 faces being named by the models trained by them. Although this test account for the ability of the models to identify faces correctly in a frame, it does not account for False Positive data, which is done in the Negative Face Data performance metric.

4.2.4 Expression Recognition Detection Speed

One of the primary aspect of Micro-Emotion Detection is that the speed of detection be fast. In particular, we need to test whether a model can identify the subtlety in Expression changes within a short time frametime. We applied a temporal frame-time analysis to understand the difference in Subtle Expression response time. Our approach is as follows:

- Create a video that cycles through the model created using FaceNet, VGGFace and MEDNet
- Detect changes in Frametime while cycling through the emotions and find avg and max frametime

Here are our results:

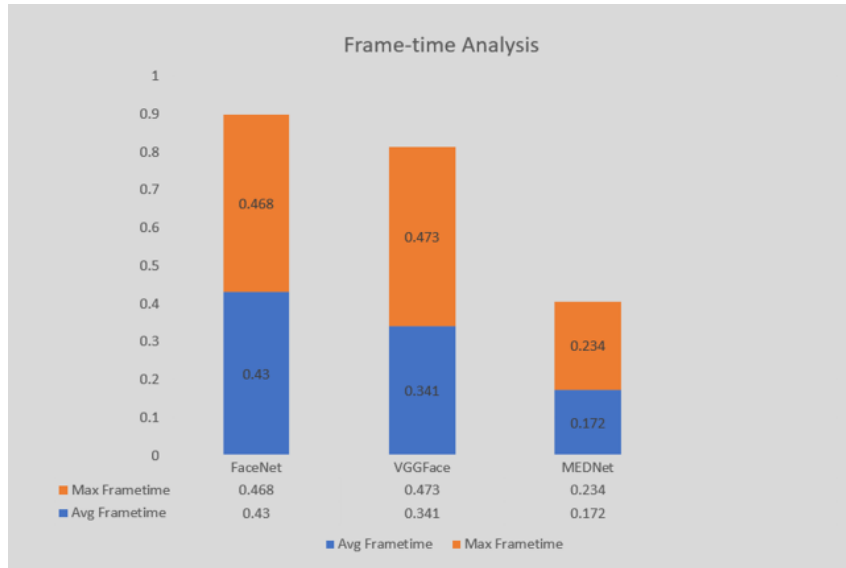


Figure 4.13: Average and Maximum Frametime

4.3 Test System Analysis

The task of identifying the Facial Features, the Classifiers and the Mediapipe implementation can be very computationally unforgiving on the system. We highly recommend a quad-core setup that supports hyperthreading, 16 GB RAM so that the buffer doesn't fill up during training and A GPU is a must for training and testing. Our test system configuration are as follows:

- CPU - core i5-11300H (Used for parallel processing and live feed analysis)
- GPU - Nvidia RTX 3060 laptop gpu (Used for Training, Testing and Opencv implementation)
- 16 GB RAM.

We recommend a series of GPU with dedicated tensor cores to perform the testing and training of the datasets.

4.4 Performance Analysis in Real-Time and Video

The LBP-CNN shows very similar results in similar results in live feed and video[20]. A mean average in accuracy of **97.9** percent was noted with a standard deviation range of **96.9 percent to 100 percent**. We tested the models accuracy on a number of students. The expressions of the students were done in both the controlled environment and through tested videos online.

- In the **Controlled Environment**, the Students were asked to read two separate paragraphs. These paragraphs started off compounding in nature, then moved on to a complete explanation of what was happening. This resulted in the students transitioning between the expressions Concentration, Confusion and Satisfaction mostly. A lot of them showed neutral expressions during

the read. It matches the average accuracy of macro-facial expression almost perfectly, sometimes performing even better. For instance, it was required that our model performs the expression recognition task very quickly due to the use of Background Subtraction and LBP implementation in the training phase. Fig 5.5 shows that the transition from Confusion to Satisfaction occurs in **1ms**. Our range extends from 1ms to upto 200ms The test paragraph read

Curiosity and excitement consumed the villagers as they clamored to experience the wonders of time travel. Ezekiel, with a mischievous grin, selected a small group of volunteers and instructed them on the machine's operation. He warned them about the risks and uncertainties of time travel, but they were undeterred.

The chosen few stepped into the time machine, and with a flick of a switch, they vanished from sight. Seconds turned into minutes, and minutes into hours, but there was no sign of their return. Panic gripped the remaining villagers as they feared the worst. Had Ezekiel sent them to their doom?

Just as despair settled in, the time machine reactivated, and one by one, the volunteers returned. Their faces were filled with awe and disbelief as they recounted their adventures. They had traveled to various points in history, witnessing pivotal events and interacting with figures they had only read about in books.

The tales grew more fantastical with each retelling. One villager claimed to have dined with Leonardo da Vinci, while another spoke of trading jokes with Mark Twain. A young woman recounted dancing with Marie Antoinette at the Palace of Versailles, and a farmer claimed to have witnessed the signing of the Declaration of Independence.

Word of Ezekiel's time machine spread far and wide, attracting adventurers, historians, and thrill-seekers from distant lands. The village transformed into a bustling hub of time travelers, and Ezekiel became a renowned figure in scientific circles.

However, as the villagers reveled in their newfound fame and wealth, a deep sense of unease settled over Ezekiel. He realized that tampering with time had consequences, and he feared the potential damage that could be done. Determined to right the wrongs, he dismantled the time machine, vowing never to allow time travel again.

The village returned to its quiet existence, and Ezekiel faded into obscurity, his secret safe within him. The tales of the time travelers became nothing more than legends whispered around campfires. And though the villagers longed for the thrill of time travel once more, they knew deep down that some secrets were best left undisturbed.

And so, the surprising story of Ezekiel and his time machine became a cautionary tale, reminding all who heard it of the delicate balance of time and the dangers that lie within its grasp.]Once upon Ezekiel claimed to have discovered the secret to time travel. At first, the villagers dismissed his claim as another one of his wild tales, but their skepticism soon turned into awe when Ezekiel

presented them with a working time machine. It was a sleek, metallic device adorned with intricate symbols and pulsating with an otherworldly energy.

Curiosity and excitement consumed the villagers as they clamored to experience the wonders of time travel. Ezekiel, with a mischievous grin, selected a small group of volunteers and instructed them on the machine's operation. He warned them about the risks and uncertainties of time travel, but they were undeterred.

The chosen few stepped into the time machine, and with a flick of a switch, they vanished from sight. Seconds turned into minutes, and minutes into hours, but there was no sign of their return. Panic gripped the remaining villagers as they feared the worst. Had Ezekiel sent them to their doom?

Just as despair settled in, the time machine reactivated, and one by one, the volunteers returned. Their faces were filled with awe and disbelief as they recounted their adventures. They had traveled to various points in history, witnessing pivotal events and interacting with figures they had only read about in books.

The tales grew more fantastical with each retelling. One villager claimed to have dined with Leonardo da Vinci, while another spoke of trading jokes with Mark Twain. A young woman recounted dancing with Marie Antoinette at the Palace of Versailles, and a farmer claimed to have witnessed the signing of the Declaration of Independence.

Word of Ezekiel's time machine spread far and wide, attracting adventurers, historians, and thrill-seekers from distant lands. The village transformed into a bustling hub of time travelers, and Ezekiel became a renowned figure in scientific circles.

However, as the villagers reveled in their newfound fame and wealth, a deep sense of unease settled over Ezekiel. He realized that tampering with time had consequences, and he feared the potential damage that could be done. Determined to right the wrongs, he dismantled the time machine, vowing never to allow time travel again.

The village returned to its quiet existence, and Ezekiel faded into obscurity, his secret safe within him. The tales of the time travelers became nothing more than legends whispered around campfires. And though the villagers longed for the thrill of time travel once more, they knew deep down that some secrets were best left undisturbed.

And so, the surprising story of Ezekiel and his time machine became a cautionary tale, reminding all who heard it of the delicate balance of time and the dangers that lie within its grasp.

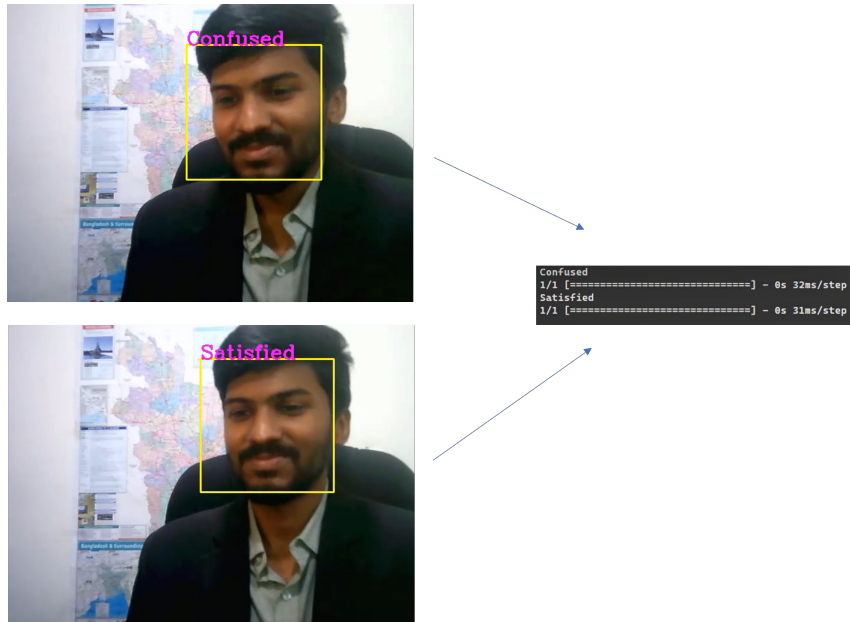


Figure 4.14: MEDNet Live Feed

It is to mention that our model showed discrepancy in facial identification when the background was busy, which greatly reduced after the implementation of LBP-CNN and showed more consistent results 2

- In case of **In the wild video feed**, the results are more of the same as well, staying well below the 250ms threshold we determined to identify micro-expressions. Fig 5.6 shows an example with mediapipe recognition implementation.

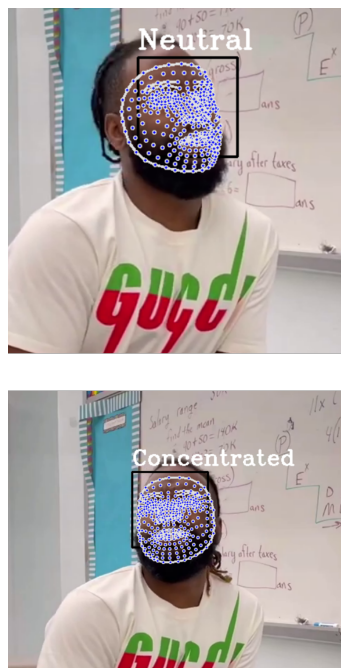


Figure 4.15: In the wild Mediapipe Implementation

Chapter 5

Conclusion

This research aimed to leverage the on-going research of Emotion Analysis. Our research will bolster the field of Facial Expression research significantly as the aim was to effectively accelerate every part of Emotion Analysis through Computer Vision, starting from training and dataset upto Real-world implementation. The results indicate that with proper tuning at every single step, Facial Expressions can be identified much more quickly and correctly.

Micro-emotion detection is a very new field of research and requires a lot more identification parameters to be maintained. Our stride here was to find an effective way of identifying unique ways to cut down the run time of the code itself. A varying number of innovative research is still left to be done. One of our goals is to implement MEDNet with a customized LBP-CNN-DT algorithm. The reason for implementing a Decision-Tree parameter will be to identify expressions before the subject produces it. For example, A person who is Neutral can show any sort of emotions, but a student who is Concentrating on his studying can either do one of the following:

- Concentration - Confusion - Concentration
- Concentration - Satisfaction - Neutral

show Confusion, back to Concentration or or a varying ways of showing the emotion. A decision-tree level can parameterize these attributes and pre-recognize them, cutting down on the speed of recognition and computational cost.

References

- [1] James A Russell and Geraldine Pratt. “A description of the affective quality attributed to environments.” In: *Journal of personality and social psychology* 38.2 (1980), p. 311.
- [2] Zhimin Zhou, Bedabrata Pain, and Eric R Fossum. “Frame-transfer CMOS active pixel sensor with pixel binning”. In: *IEEE Transactions on electron devices* 44.10 (1997), pp. 1764–1768.
- [3] Paul Ekman. “Facial expressions”. In: *Handbook of cognition and emotion* 16.301 (1999), e320.
- [4] Valery A Petrushin. “Emotion recognition in speech signal: experimental study, development, and application”. In: *Sixth international conference on spoken language processing*. 2000.
- [5] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. “Toward machine emotional intelligence: Analysis of affective physiological state”. In: *IEEE transactions on pattern analysis and machine intelligence* 23.10 (2001), pp. 1175–1191.
- [6] Navneet Dalal and Bill Triggs. “Histograms of Oriented Gradients for Human Detection”. In: vol. 1. July 2005, pp. 886–893. DOI: 10.1109/CVPR.2005.177.
- [7] MJ Den Uyl and H Van Kuilenburg. “The FaceReader: Online facial expression recognition”. In: *Proceedings of measuring behavior*. Vol. 30. 2. Citeseer. 2005, pp. 589–590.
- [8] Maja Pantic et al. “Web-based database for facial expression analysis”. In: *2005 IEEE international conference on multimedia and Expo*. IEEE. 2005, 5–pp.
- [9] Lianghua He et al. “An enhanced LBP feature based on facial expression recognition”. In: *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*. IEEE. 2006, pp. 3300–3303.
- [10] Adam Schmidt and Andrzej Kasiński. “The performance of the haar cascade classifiers applied to the face and eyes detection”. In: *Computer Recognition Systems 2*. Springer. 2007, pp. 816–823.
- [11] Hao Li et al. “Image restoration after pixel binning in image sensors”. In: *Tsinghua Science and Technology* 14.4 (2009), pp. 541–545. DOI: 10.1016/S1007-0214(09)70114-2.
- [12] Neeta Sarode and Shalini Bhatia. “Facial expression recognition”. In: *International Journal on computer science and Engineering* 2.5 (2010), pp. 1552–1557.

- [13] Yan Ma. “Number Local binary pattern: An Extended Local Binary Pattern”. In: *2011 International Conference on Wavelet Analysis and Pattern Recognition*. 2011, pp. 272–275. DOI: 10.1109/ICWAPR.2011.6014464.
- [14] Qi Wu, Xunbing Shen, and Xiaolan Fu. “The machine knows what you are hiding: an automatic micro-expression recognition system”. In: *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*. Springer. 2011, pp. 152–162.
- [15] DK Kirange and RR Deshmukh. “Emotion classification of news headlines using SVM”. In: *Asian Journal of Computer Science and Information Technology* 5.2 (2012), pp. 104–106.
- [16] Pushpaja V Saudagare, DS Chaudhari, et al. “Facial expression recognition using neural network—An overview”. In: *International Journal of Soft Computing and Engineering (IJSCIE)* 2.1 (2012), pp. 224–227.
- [17] Ian Goodfellow et al. *Challenges in Representation Learning: A report on three machine learning contests*. 2013. URL: <http://arxiv.org/abs/1307.0414>.
- [18] Xiaobai Li et al. “A spontaneous micro-expression database: Inducement, collection and baseline”. In: *2013 10th IEEE International Conference and Workshops on Automatic face and gesture recognition (fg)*. IEEE. 2013, pp. 1–6.
- [19] Daniel McDuff et al. “Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2013, pp. 881–888.
- [20] Matthew Shreve. *Automatic macro-and micro-facial expression spotting and applications*. University of South Florida, 2013.
- [21] Wen-Jing Yan et al. “For micro-expression recognition: Database and suggestions”. In: *Neurocomputing* 136 (2014), pp. 82–87.
- [22] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu. “Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset”. In: *IEEE Transactions on Multimedia* 17.6 (2015), pp. 804–815.
- [23] Chen-Chiung Hsieh et al. “Effective semantic features for facial expressions recognition using SVM”. In: *Multimedia Tools and Applications* 75 (Apr. 2015). DOI: 10.1007/s11042-015-2598-1.
- [24] Ziwei Liu et al. “Deep Learning Face Attributes in the Wild”. In: *Proceedings of International Conference on Computer Vision (ICCV)*. Dec. 2015.
- [25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.
- [26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.
- [27] Amir Mohammad Shahsavarani et al. “Assessment & measurement of anger in behavioral and social sciences: a systematic review of literature”. In: *International Journal of Medical Reviews* 2.3 (2015), pp. 279–286.

- [28] Xincheng Ye et al. “Foreground–background separation from video clips via motion-assisted matrix restoration”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 25.11 (2015), pp. 1721–1734.
- [29] Anas Abouyahya et al. “Features extraction for facial expressions recognition”. In: *2016 5th International Conference on Multimedia Computing and Systems (ICMCS)*. IEEE. 2016, pp. 46–49.
- [30] Neha Bhardwaj and Manish Dixit. “A review: facial expression detection with its techniques and application”. In: *International Journal of Signal Processing, Image Processing and Pattern Recognition* 9.6 (2016), pp. 149–158.
- [31] Xiaohong Li, Jun Yu, and Shu Zhan. “Spontaneous facial micro-expression detection based on deep learning”. In: *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. 2016, pp. 1130–1134. DOI: 10.1109/ICSP.2016.7878004.
- [32] SG Mangalagowri and P Cyril Prasanna Raj. “EEG feature extraction and classification using feed forward backpropagation algorithm for emotion detection”. In: *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*. IEEE. 2016, pp. 183–187.
- [33] Rudrika Kalsotra and Sakshi Arora. “Morphological based moving object detection with background subtraction method”. In: *2017 4th International Conference on Signal Processing, Computing and Control (ISPCC)*. 2017, pp. 305–310. DOI: 10.1109/ISPCC.2017.8269694.
- [34] Eva G Krumhuber et al. “A review of dynamic datasets for facial expression research”. In: *Emotion Review* 9.3 (2017), pp. 280–292.
- [35] Shan Li, Weihong Deng, and JunPing Du. “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017, pp. 2584–2593.
- [36] Shunji Mitsuyoshi et al. “Mental status assessment of disaster relief personnel by vocal affect display based on voice emotion recognition”. In: *Disaster and military medicine* 3 (2017), pp. 1–9.
- [37] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. “Affectnet: A database for facial expression, valence, and arousal computing in the wild”. In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.
- [38] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.
- [39] Madhumita A. Takalkar and Min Xu. “Image Based Facial Micro-Expression Recognition Using Deep Learning on Small Datasets”. In: *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 2017, pp. 1–7. DOI: 10.1109/DICTA.2017.8227443.
- [40] Jiajiong Cao, Yingming Li, and Zhongfei Zhang. “Celeb-500k: A large training dataset for face recognition”. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 2406–2410.

- [41] Dong Yoon Choi, Dae Ha Kim, and Byung Cheol Song. “Recognizing fine facial micro-expressions using two-dimensional landmark feature”. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2018, pp. 1962–1966.
- [42] Jiachao Zhang et al. “Pixel binning for high dynamic range color image sensor using square sampling lattice”. In: *IEEE Transactions on Image Processing* 27.5 (2018), pp. 2229–2241.
- [43] Shan Li and Weihong Deng. “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition”. In: *IEEE Transactions on Image Processing* 28.1 (2019), pp. 356–370.
- [44] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild”. In: *IEEE Transactions on Affective Computing* 10.1 (2019), pp. 18–31. DOI: 10.1109/TAFFC.2017.2740923.
- [45] Jiayin Pei and Peng Shan. “A Micro-expression Recognition Algorithm for Students in Classroom Learning Based on Convolutional Neural Network.” In: *Traitement du Signal* 36.6 (2019).
- [46] Feiyu Xu et al. “Explainable AI: A brief survey on history, research areas, approaches and challenges”. In: *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II* 8. Springer. 2019, pp. 563–574.
- [47] Shan Li and Weihong Deng. “Deep facial expression recognition: A survey”. In: *IEEE transactions on affective computing* 13.3 (2020), pp. 1195–1215.
- [48] M.A. NASRI et al. “Face Emotion Recognition From Static Image Based on Convolution Neural Networks”. In: *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. 2020, pp. 1–6. DOI: 10.1109/ATSIP49331.2020.9231537.
- [49] Lutfiah Zahara et al. “The facial emotion recognition (FER-2013) dataset for prediction system of micro-expressions face using the convolutional neural network (CNN) algorithm based Raspberry Pi”. In: *2020 Fifth international conference on informatics and computing (ICIC)*. IEEE. 2020, pp. 1–9.
- [50] P Mary Jenifer et al. “Multiple Face Detection and Attendance System Using OpenCV”. In: *2021 International Conference on Simulation, Automation Smart Manufacturing (SASM)*. 2021, pp. 1–5. DOI: 10.1109/SASM51857.2021.9841223.
- [51] M Murugappan. *FEER Dataset*. 2021. DOI: 10.34740/KAGGLE/DSV/1856006. URL: <https://www.kaggle.com/dsv/1856006>.
- [52] Chunming Wu and Ying Zhang. “MTCNN and FACENET based access control system for face detection and recognition”. In: *Automatic Control and Computer Sciences* 55 (2021), pp. 102–112.
- [53] Hasan Zan and Abdulnasır Yıldız. “Sleep Arousal Detection Using One Dimensional Local Binary Pattern-Based Convolutional Neural Network”. In: *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. 2021, pp. 1–4. DOI: 10.1109/INISTA52262.2021.9548369.

- [54] Aman and AL Sangal. “Drowsy Alarm System Based on Face Landmarks Detection Using MediaPipe FaceMesh”. In: *Proceedings of First International Conference on Computational Electronics for Wireless Communications: IC-CWC 2021*. Springer. 2022, pp. 363–375.
- [55] Tashreef Abdullah Araf et al. “Real-Time Face Emotion Recognition and Visualization using Grad-CAM”. In: *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. 2022, pp. 1–5. DOI: 10.1109/ICAECT54875.2022.9807868.
- [56] Jingting Li et al. “FME’22: 2nd Workshop on Facial Micro-Expression: Advanced Techniques for Multi-Modal Facial Expression Analysis”. In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 7397–7399.
- [57] Yadira Quiñonez, Carmen Lizarraga, and Raquel Aguayo. “Machine Learning solutions with MediaPipe”. In: *2022 11th International Conference On Software Process Improvement (CIMPS)*. 2022, pp. 212–215. DOI: 10.1109/CIMPS57786.2022.10035706.
- [58] Andrey V Savchenko. “Frame-level prediction of facial expressions, valence, arousal and action units for mobile devices”. In: *arXiv preprint arXiv:2203.13436* (2022).
- [59] B. Thaman, T. Cao, and N. Caporusso. “Face Mask Detection using MediaPipe Facemesh”. In: *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*. 2022, pp. 378–382. DOI: 10.23919/MIPRO55190.2022.9803531.
- [60] Andrian Firmansyah, Tien Fabrianti Kusumasari, and Ekky Novriza Alam. “Comparison of Face Recognition Accuracy of ArcFace, Facenet and Facenet512 Models on Deepface Framework”. In: *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*. 2023, pp. 535–539. DOI: 10.1109/ICCoSITE57641.2023.10127799.
- [61] Sakthimohan M et al. “Detection and Recognition of Face Using Deep Learning”. In: *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*. 2023, pp. 72–76. DOI: 10.1109/ICISCoIS56541.2023.10100435.
- [62] M Monica Dhana Ranjini et al. “Haar Cascade Classifier-based Real-Time Face Recognition and Face Detection”. In: *2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC)*. 2023, pp. 990–995. DOI: 10.1109/ICESC57686.2023.10192586.