National College of Ireland

A Comparative Study of Pixel Values and Landmark Detection Features to Solve for Facial Emotion Recognition

MSc Research Project
Data Analytics

Tiago Leonel do Nascimento
Student ID: 19143486

School of Computing
National College of Ireland

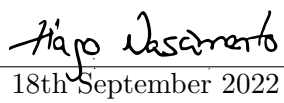
Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Tiago Leonel do Nascimento
Student ID:	19143486
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	15/08/2022
Project Title:	A Comparative Study of Pixel Values and Landmark Detection Features to Solve for Facial Emotion Recognition
Word Count:	7348
Page Count:	26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	18th September 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Comparative Study of Pixel Values and Landmark Detection Features to Solve for Facial Emotion Recognition

Tiago Leonel do Nascimento
19143486

Abstract

Understanding and recognizing various human emotions is foundational to how our society works. In recent decades, researchers have tried to train machine learning models that can replicate emotion recognition by utilizing Facial Emotion detection. This study aims to build different machine learning models that can effectively classify human emotions using Facial Emotion Recognition (FER) and produce a comparison between pixel values and landmark detection as features to achieve the classification. Using both posed and spontaneous datasets, different models such as Random Forest, Extra Trees classifier, Support Vector Machine and a Convolutional Neural Network are developed to achieve high accuracy. An ensemble machine learning model using hard voting created utilizing Random Forest, Extra Trees Classifier and SVM with grid search achieved 89% accuracy using pixel values as features.

1 Introduction

Emotion is a complex topic explored in various domains, including psychology and computer science. Understanding and recognizing various human emotions is a foundational component of how our society works and an essential factor in human daily experiences and survival. Humans have the inherent ability to interpret and understand emotions, and in recent decades researchers have tried to train machine learning models that can replicate emotion recognition. Identifying a friendly face, anger, or a dangerous threat has helped our society evolve (Comas et al.; 2020; Haghpanah et al.; 2022; Ong; 2021; Stark and Hoey; 2021). Rosalind Picard (Aranha et al.; 2021; Wessler et al.; 2021) coined the term Affective Computing (AC) which looks into the recognition, interpretation, processing, and simulation of human emotions. Facial Expression Recognition (FER) is the field that looks into facial expressions to understand facial identification and emotional state. The advancements in emotion recognition and classification have been shaped by many computer capability advancements and deep learning developments in the recent years(Alqahtani et al.; 2019; Booth et al.; 2021; Haghpanah et al.; 2022; Ong; 2021; Stark and Hoey; 2021). This study aims to build different machine learning models that can effectively classify human emotions.

1.1 Motivation

As a whole, emotion interpretation can be subjective in human-to-human exchanges, and much of how we interpret emotions is influenced by our surroundings and past experiences. Different theories of emotion showcase different approaches used to try to explain emotions. Although researchers frequently utilize emotion theories, there is no consensus on how accurate they are in practice, how inclusive they can be of diverse cultures, or how to improve them. While there is no global agreement in the scientific community on what emotions are and how they work, different AC models and commercial solutions are deployed, raising concerns around ethics and bias (Stark and Hoey; 2021).

Researchers creating solutions for FER face challenges of technical nature and navigating the uncertainty of what emotions are. Moravec's paradox[1], which explains the difficulties in teaching computers to execute activities that humans consider simple, is one of the essential concepts in Artificial Intelligence (AI). The paradox explains that while some human abilities, such as motor skills and the ability to generalize, are second nature, they are complicated for computers to perform. Meanwhile, tasks that are difficult for humans, such as extensive math computation and logic calculations, are standard tasks for computers.

One of the technical issues faced by researchers in AC, as explored by Takalkar and Xu (2017) is issues around accuracy in labelling emotions, with some emotional classes producing less accurate results than others. According to the author, some expressions are so nuanced that even humans have difficulty determining the correct label for a depicted emotion during the validation phase of an FER database creation. Different people have different reactions to the same pre-recorded emotion. As machine learning models learn from humans, bias can affect models; as humans struggle to identify emotions and their definitions, so do the models that we create Masson et al. (2020); Booth et al. (2021).

1.2 Research Question and Objectives

The research questions that will be explored in this paper are:

1. How does the usage of traditional machine learning techniques compare to deep learning solutions' accuracy?

2. Does the usage of Landmark detection points increase model accuracy in comparison with pixel values?

3. Which baseline or fine-tuned models outperform others in classifying Facial Emotion Recognition classes?

Based on the explored literature review in the related works section, this work will compare pixel value features and facial landmark detection used to train machine learning (ML) and a Convolutional Neural Networks (CNN) model. The main contributions of this paper are:

- Creation of multiple algorithms that classify FER using different methods such as K-nearest neighbor (KNN), Random Forest, Extra Trees classifier, Support Vector Machine (SVM), Ensemble ML model and a CNN.

[1]Moravec's paradox: https://medium.com/@froger_mcs/moravecs-paradox-c79bf638103f

- Performance comparison between ML models in both posed and spontaneous emotion datasets.
- Comparison of accuracy between single and multicultural datasets.

1.3 Paper Structure

This paper is further divided into different sections. Section 2 looks into related works and discusses relevant papers from 2017 to 2022, looking at emotion theory models, FER, and landmark detection. Section 3, research methodology, explores the different approaches for dataset selection, pre-processing and algorithms used. Section 4 covers information on design specifications and implementation, providing the framework utilised to reach the solution proposed by this paper and discusses the outputs of the project, including code written and models' development. Section 5, evaluation, looks into the results collected by this research and the overall models' performance. Section 6, conclusion and future work, closes the paper with the findings and how successfully the research question is answered. Future work guides how future studies could further the research this paper has started and what limitations impacted the final results.

2 Related Work

This section provides a literature review on recent publications in Facial Expressions Recognition. The Related Work section comprises peer-reviewed studies published between 2017 and 2022 discussing FER and emotional classification. The first part delves into some of the most used types of emotion theory researchers employ to drive the creation of databases, FER models and emotion classification. The second part of the related works looks into different primary papers and the techniques used to create FER solutions. A comparison of features utilized, techniques and achieved accuracy is produced, and a brief conclusion closes this study section.

2.1 Emotion Identification, Commonly agreed research methods

The paper titled by Masson et al. (2020) is a review of 220 papers on emotion identification. The author describes how different academics employ emotional models to identify emotions and stresses the difficulty of constructing models to address AC. The following are some of the most common types of emotional model theory groups used in emotional identification and AC research:

- Basic Emotions
- Action Units
- Complex Emotions model
- Valence and Arousal

Basic Emotions was theorised by Paul Ekman and Carrol E. Izard (Masson et al.; 2020; Stark and Hoey; 2021) is the most widely accepted emotion theory, utilised in 70%

of the 220 papers reviewed by Masson et al. (2020). It is based on Darwin's original ideas that some emotions are hardwired as second nature to humans. These raw emotions (fear, surprise, sadness, joy, anger and disgust) support us in adapting to our environment and social interactions. As per the theory, these emotions are observed in humans across all cultures with little to no variance. The Action Units (AUs) theory investigates the movement of various facial muscles and classifies their movement as a response to various emotions. Complex Emotion Theory looks into combinations, usually two or more, emotions formed from basic emotions. The Valence and Arousal theory views emotions as a continuous organisation rather than a distinct entity. Arousal examines the intensity of the exhibited emotion, whereas valence examines the emotional scale, which can be positive or negative. Instead of using a discrete, categorical list model for valence and arousal, emotions can be plotted in a two- or three-dimensional space (Masson et al.; 2020).

While FER researchers widely accept the theory of the basic emotions as a means of identifying them, other academics have expressed concerns about how the ideas are not challenged and no additional research is done to understand better how we deal with emotion classification (Stark and Hoey; 2021; Masson et al.; 2020). According to Masson et al. (2020), the current theoretical and technical debates and lack of agreement make the approaches to solve for AC convoluted. The author also believes that the categorical approach of the basic emotions model is being phased out in favour of more robust theoretical models that better account for the complexity of emotions.

Models that identify human emotion can be created using a variety of methods. One of these methods is Landmark detection. Landmark detection is a computer vision task that can map facial features to different key points around a subject's face. Those points can vary in number and usually are positioned around eyes, eyebrows, nose, mouth and face contouring. These points can be used directly to infer emotions based on their position on a face or distance. Landmark detection first identifies a subject's face and then applies points to specific areas. Landmark detection applications usually focus on two approaches: mapping the face from a neutral expression until peak emotion is reached or utilizing a set of static images where the subject is displaying an emotion. (Haghpanah et al.; 2022; Qiu and Wan; 2019; Özseven and Düğenci; 2017). This work focuses on pixel values and landmark detection features, which are collected from static images, as a way to solve FER problems. Pixel Values look into the intensity, brightness and respective color of a pixel with numbers in a range from 0 (black) to 255 (white). Facial landmark detection was chosen as it identifies facial features and highlights them; This is something the researcher initially assumed to be very important, particularly for spontaneous datasets, where different backgrounds, poses and artefacts could potentially impact to the accuracy results from pixel values alone. The assumption is that landmark detection and distances could increase accuracy by keeping track of the points while pixel values does not have this precise mapping of facial features and captures pixel values for the entire image.

Other methods explored to solve for AC and emotion classification are the use micro-expressions, where computer vision is used to look for micro-movements in human facial expressions (Wang et al.; 2020; Stanciu and Albu; 2019), Electrodermal Activity (EDA) sensors (Canabal et al.; 2020), Electroencephalography (EEG) (Shen et al.; 2020), electrocardiogram (ECG) (Alqahtani et al.; 2019), audio files with vocal recordings, infrared sensors and changes in semantics (Wang et al.; 2021).

The accuracy of the created models varies depending on the method selected. While

Picard's concepts for Affective Computing have been widely accepted, other academics have recommended that emotion classification employ a broader set of methods rather than only tracking and sensing computational data.

2.2 Usage of Landmark Detection and Pixel Values in Facial Emotion Recognition

This section discusses papers with relevant usage of pixel values, landmark detection, and FER applications. The papers selected make different decisions, both on how features are used and what machine learning or deep learning models are developed to achieve emotion classification.

The paper by Özseven and Düğenci (2017) uses distance and slope between 14 different landmark detection points. A dataset named BioID composed of 1523 images and 23 subjects was used in this study. Different ML techniques are also explored in this paper. The author states that the higher the number of points used in landmark detection, the higher the accuracy is observed, mentioning works that utilize 68 points; however, the author decides to focus on 14 points. Although the number of points extracted from the different images is smaller than other mentioned works in this literature review, the author reports an accuracy of 94.6% utilizing an MLP classifier model.

The study by Barman and Dutta (2017), published in the 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN) makes use of an MLP and Nonlinear AutoRegressive with eXogenous (NARX) model and landmark detection to classify emotion using the datasets JAFFE, CK+, MUG and MMI. Landmarks are extracted, and a grid is created to be used as features to train the models. For the JAFFE dataset, 92.0% accuracy is achieved utilizing an MLP model and 97.6% for NARX. Results for CK+ reach 100% accuracy using NARX.

Paper by Thannoon et al. (2018) makes use of 8 selected AUs and utilizes these points to classify emotion deception using MLP, VG-RAM, Support Vector Machine (SVM) and k-nearest neighbors (KNN), with KNN and VG-RAM producing the best accuracy. Although not trying to identify basic emotions or other sets of emotions like the rest of the papers cited in this work, it is still relevant based on the applied methodology and process used by the author to reach the results. This paper separated results based on genders achieving the highest accuracy at 90% for males in both KNN and VG-RAM and 84% for both genders. Results are compared to different features, with the 8 AUs exceeding accuracy for most of the results apart from Micro-expressions at 85%.

Author Wu and Ji (2019) discusses using datasets created in near-perfect conditions with less challenging images. These datasets usually tend to have perfect lighting and background and receive a much higher accuracy, but this accuracy does not translate when models are used in real-world scenarios. The author mentions a shift in interest from researchers in the past decade, from posed datasets to datasets containing "in-the-wild" conditions. "In-the-wild" datasets offer a more diverse set of images, subjects facial features, and artefacts, for example, having spontaneous poses and artefacts such as sunglasses, different beard styles that could partially cover a subject's face, hats and others.

The article published in the 2019 International Conference on Advanced Science and Engineering (ICOASE) by the author Dino and Abdulrazzaq (2019), utilizes machine learning techniques and MLP to classify eight emotions utilizing the dataset CK+. The author resized the images to 256 by 256 pixels in greyscale, and a PCA is also used for

feature dimensional reduction and suggests that other researchers use PCA to reduce redundant feature information. The 10-fold Cross-validation method is used for the validation of results, with SVM achieving the highest accuracy score at 93.89%, followed by MLP neural network at 82.97%. The author shared a table with information on the number of detected faces per class which seems to indicate an imbalance in the number of examples used, such as 50% neutral examples (317 examples) against 4.1% representing 26 examples for the emotion sadness.

Author Qiu and Wan (2019) utilises 68 landmark detection points and the CK+ dataset to classify seven emotions, stating that reported accuracy using traditional ML is comparable to state-of-the-art CNN-based models. In this research, points are distributed in the subject's face and normalised based on the point in the subject's nose or on multiple points named origin points. Classification is achieved by utilising an MLP with three layers. The proposed approach reported for one-origin points is 87% and 92% for multi-origin points. The paper also produced a running time comparison of different state-of-the-art models such as VGG16(Simonyan and Zisserman; 2014)and Resnet18(He et al.; 2016), showing it produced results in a much faster manner. Not all papers reported on time to run models so this information is only discussed above.

Active Shape Modeling (ASM) is used in this study. The author Cao and Qi (2021) proposes an algorithm that can construct a three-dimensional face model with 48 face feature points and feature vectors composed of 10 connecting lines to recognise different emotions (angry, tired, daze, pleasure, sad, interest, normal) reaching an accuracy of 69%. This type of tracking is meant to be utilised in a classroom environment to understand students' attention and emotions. This is a fascinating study because it looks at classifying emotions outside the usual six basic emotions but looks at FER specifically from an interesting angle for a classroom environment.

The study by Chouhayebi et al. (2021) utilised three different datasets to classify emotions: happiness, surprise and neutrality. The author utilised two different algorithms to perform the classification, SVM reaching 91.5% accuracy and MLP at 96%. The author extracted features utilising different methods, including features collected individually from different parts of the subject face and the usage of landmark detection. The dlib library and 68 landmark points were collected. The author states that for smaller datasets, SVM performs better, whereas MLP achieves better accuracy for larger datasets. Results on the different experiments with a smaller dataset achieved over 90% accuracy, with dlib landmark detection achieving 98.5%.

The paper by Maurya and Sharma (2022) created a CNN for classifying emotions from images. CNN is utilised as it can automatically detect essential features for image classification and emotion recognition without needing supervision/feeding the correct features. The author used images of 48 by 48 pixels and five emotional classes. Accuracy reported after training is around 69%

The paper by Haghpanah et al. (2022) discusses the application of 68 landmark detection points and the Facial Action Coding System (FACS) in a neural network model to identify human emotions in the Cohn-Kanade (CK+) dataset. Landmarks are collected utilising the python library dlib (King; 2009). In this paper, the features are extracted and used in a single neural network. The author calls out that the euclidean distance is used as a feature in a mentioned paper in their literature review. This study will also use the euclidean distance between points similar to this paper; however, it will also be applied in different datasets and compared with other features. The author uses a considerably small number of images to train their model, but MLP enables them to achieve

96% accuracy in their test data. This MLP model reached 2-3% lower results than the state-of-the-art CNN-based models.

The author Cho et al. (2022) introduces a novel multi-label dataset with over 38 thousand images classifying Korean drama video clips into 23 different emotions. The author went through different categorisation and labelling processes and utilised Autoencoder, CNN and transfer learning. Images were classified manually, and later a model was created to classify the new dataset. The dataset is used with ResNet50 achieving 68.03% accuracy.

2.3 Comparison of Landmark Detection Reviewed Results

Table 1 provides an overview of the results discussed in the section above. Results from the different studies vary in accuracy, achieving between 68-100% reported accuracy. Most published papers utilised 68 facial landmark points and a mixture of machine learning techniques, with KNN and SVM, highlighted.

Table 1: Features and Techniques presented in the Literature Review.

Author Names	Feature / Technique	Accuracy
Özseven and Düğenci (2017)	14 points. MLP	94.6%
Barman and Dutta (2017)	Salient Landmarks. MLP, NARX	92% - 100%
Thannoon et al. (2018)	8 AUs. MLP, VG-RAM, SVM, KNN	84% - 90%
Cao and Qi (2021)	48 points, 10 connecting lines. ASM	69%
Dino and Abdulrazzaq (2019)	SVM, KNN, MLP	82.97% - 93.89%
Qiu and Wan (2019)	68 points. MLP	87% - 92%8
Maurya and Sharma (2022)	CNN extracted features	69%
Cho et al. (2022)	Resnet	68.03%
Chouhayebi et al. (2021)	68 points	91.5% - 96%
Haghpanah et al. (2022)	68 points. FACs. MLP	96%

2.4 Related Works Conclusion

This section covered current discussions in the research community about emotion recognition and some of the different options utilized to solve emotion classification issues. The Usage of Landmark Detection in Facial Emotion Recognition section covered papers that utilized landmark detection and different machine learning techniques, reaching accuracy comparable with state-of-the-art models while being faster to train results. SVM and KNN are mentioned in various papers with good results, so these techniques will be utilized in this research. Dlip library will also be utilized to extract facial landmark point features.

3 Methodology

3.1 Introduction

This section will focus on the Methodology used, which takes inspiration from the cross-industry process for data mining (CRISP-DM) methodology (Wirth and Hipp; 2000) and the selected approaches to feature extraction and classification present in the literature review, discussed in the previous section. The methodology section looks into Business Understanding, Data Understanding, Data-preparation, Modeling, and Evaluation. Each stage of the process is explained below, and figure 1 shows a graphic representation of CRISP-DM.

CRISP-DM Methodology Applied for Image Feature Classification

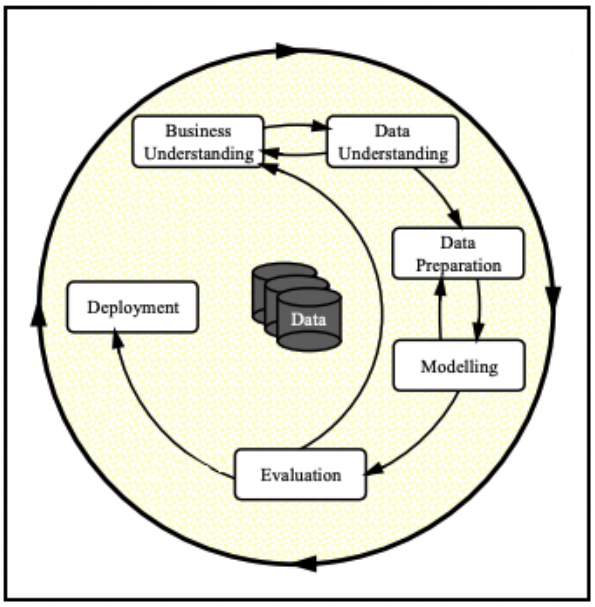


Figure 1: Cross-industry process for data mining (CRISP-DM) (Wirth and Hipp; 2000)

3.2 Business Understanding

As we evolve technically and as a society, the applications of FER and emotion classification gain traction and advancements. The literature review shows that FER can obtain significant results even when using traditional machine learning. Achieving good classification results with models that can produce good accuracy at short running times and resources opens up the possibilities for multiple researchers and companies to create FER solutions. For businesses, understanding a consumer reaction to a product, ad, or customer support received or tracking emotions from how a customer arrives to seek solutions and what emotion they display after having their issue addressed might prove very valuable. In psychology, recognising an individual's state of mind and emotion and tracking how they respond to stimuli and progress over time is another application of FER and emotion classification. The same tools and systems can also be used for image recognition and are utilised by different governments and police from different countries

such as China[2], [3] and the United Kingdom[4]. There is an ongoing discussion of AI and FER's impacts on our society with both positive and negative outlooks. Researchers must be aware of bias issues and mindful of the solutions they create.

3.3 Data Understanding

The literature review guided the research to answer the research questions, and based on the papers added as references; different datasets were selected as possible candidates to complete this study. Datasets are built differently and to suit different purposes. Datasets can be posed, spontaneous or mixed, and their images offer different combinations of subjects which can be male, female, young or old, presenting many different physical features and image backgrounds. Some datasets aim to aid in classifying 2 emotions (smiling or not smiling), five emotions, and seven emotions, such as basic emotions and complex combinations resulting in up to 18 emotion classes.

The selected datasets are available under request and EULA signing. Different datasets had different waiting times to receive a response and link to download the data. Most had the requirement that a faculty professor request access on the student's behalf.

1. The posed dataset Japanese Female Facial Expression (JAFFE) created by Lyons et al. (2020) consists of 10 Japanese female subjects displaying the 6 basic facial emotions plus neutral emotion. There are 213 images in total in 256 by 256 pixels grey-scale. Images are in the Tiff format with no compression. Images were created with excellent lighting and background. This dataset was created over 25 years and is utilized in many different papers (Lyons; 2021).

2. The spontaneous dataset Static Facial Expressions in the Wild (SFEW) (Dhall et al.; 2012) was selected to contrast the JAFFE dataset for having "in-the-wild" type of images. The dataset contains 700 images and 68 subjects. The images were collected from different movies and display a variety of genders, occlusions, different illumination and poses, being referred as close to real life environment by the authors.

The difficulty of classifying emotions increases with SFEW as the near-perfect to perfect conditions found on the JAFFE dataset are erased in SFEW_2.0. Training accuracy is impacted, but the results are closer to the ones achieved in real-life usage scenarios. Masson et al. (2020) states that a clear preference can be seen from researchers in choosing datasets for Affecting Computing where 74% of researchers part of their 220 paper reviews, chose posed datasets over spontaneous and 29% of the 220 studies worked with the JAFFE dataset. Accuracy in the different papers has been increasing with the author mentioning an average overall accuracy increase from 83.5% to 85.3% between 2014 and 2019 papers. There are considerable differences in accuracy between posed and spontaneous results, as reported in this study. Masson et al. (2020) reports that the average accuracy of posed datasets was 21.29 points higher than for spontaneous datasets, with the average accuracy of spontaneous datasets at 63.75% in their study.

[2]AI emotion-detection software tested on Uyghurs:https://www.bbc.com/news/technology-57101248

[3]China growing use of emotion recognition tech raises rights concerns: https://www.reuters.com/article/china-growing-use-of-emotion-recognition-idUSL8N2K2500

[4]London is buying heaps of facial recognition tech: https://www.wired.co.uk/article/met-police-facial-recognition-new

3.4 Data Preparation

In this step, data is preprocessed. The selected datasets were stored in Google Drive for easy access. The researcher utilised Google Colaboratory (Colab) to conduct the research. Images are retrieved utilising OpenCV library, reading from their existing folders. During the image importing, labels were created by appending their values based on their stored folder names. Images were resized to 224 by 224 pixels and converted from GBR (Green, Blue, Red) to the more traditional RGB (Red, Green and Blue). Here data visualisation takes place by printing the different types of emotion being displayed.

From the loaded data, three different sets of features were extracted. Pixel data both in RGB and greyscale to suit the different datasets. Pixel data is also normalised by dividing the collected values by 255, the maximum value for pixels. Sixty-eight landmark detection points were extracted utilising the library Mlxtend. The number of points were chosen based on the literature review (Haghpanah et al.; 2022; Chouhayebi et al.; 2021; Qiu and Wan; 2019). Calculating the euclidean distance between landmark points is also used to infer emotion.

For landmark detection, Mlxtend presented difficulties finding faces in most of the "in-the-wild" pictures. This resulted in a much smaller number of available images to train models using landmark detection—more details in the implementation section.

3.5 Modeling

The algorithms selected to be part of this research were guided by the literature review and domain research to answer the research questions. Although many recent publications in the past decades have transitioned from utilizing traditional ML to deep learning, papers in FER, particularly those utilizing Landmark Detection, rely on traditional ML. This could be due to the results achieved utilizing traditional machine learning, which is comparable to state-of-the-art deep learning models and can be processed more quickly.

The models selected will aid in answering the research question. KNN, Random Forest, Extra Trees, SVM and Ensemble of highest scoring models are selected for results comparison. Pixel value features will be used to train both ML algorithms and a CNN model. Landmark detection and landmark distances will be used with ML techniques.

Further information is discussed on models and model architecture in the next section of the research paper.

3.6 Model Evaluation

This study utilises different approaches for model evaluation. The selected methods are splitting the data into Train, Test and Train, and Test and Validation for the CNN. Stratify is also used to ensure balance amongst the different classes.

A Confusion Matrix was also utilised to evaluate the different performances of the models and the different emotion scoring across different models and datasets.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN)$$

4 Design Specification

The techniques and framework utilized in this research are explored in depth in this section. As mentioned, datasets are created differently and can have different angles, subjects of a singular gender or males and females. They can be representative of various ethnicities or just one. For this reason, results will be compared with the accuracy achieved when applying the same algorithms but not between datasets. Equally, the type of features extracted plays an essential role in some aspects of the framework used to classify emotions. Figure 2 displays the process flow framework followed to aid in completing this study and answering the research questions. In preprocessing, the images are read using OpenCV, converted from BGR to RGB and resized to 224 by 224. The labels are created, and data augmentation is considered. In Transformation, different features are extracted, both Landmark points and distances and Pixel values, missing values removed, and the data is split into train and testing and ready for modelling. Different ML techniques are applied in Modeling, and model evaluation takes place. Further details are explained in the following sections.

Framework for FER Image Feature Classification

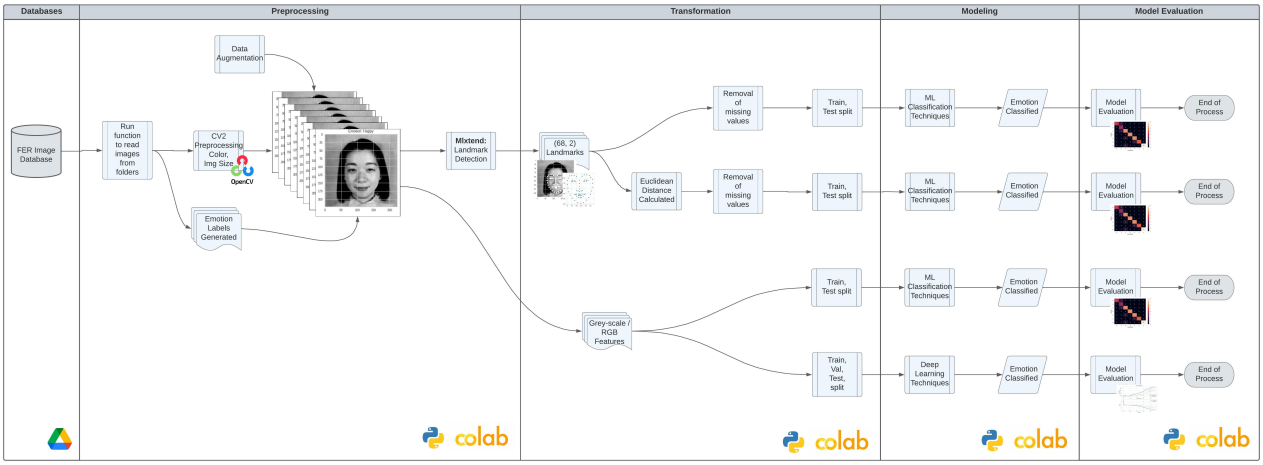


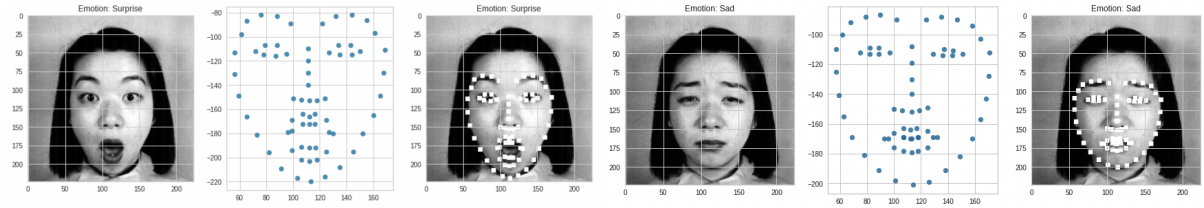
Figure 2: Structure for Image Facial Emotion Recognition

4.1 Data Augmentation

Data augmentation was attempted to increase the number of available samples. JAFFE and SFEW have a reasonable number of examples to train a machine learning model but not enough to train models using deep learning architecture. Data augmentation was applied at some stages of the project but not all, creating new images from existing ones by applying random zoom and random rotation to images. This type of data augmentation was used in the CNN creation with JAFFE. A different type of augmentation was applied to SFEW images utilizing the python library Augmentor, applying distortion, flipping images from left to right, rotating and applying a random crop.

4.2 Grey-Scale and RGB Pixel Value Feature

Grey-Scale and RGB Pixel Value Features are extracted from the images and flattened to serve as input for the different ML models created. Different stratified data splits are cre-



(a) Surprise (b) Sadness

Figure 3: Surprise and Sadness. Faces with landmark detection points.

ated as a result of feature extraction, PCA, LDA and data normalisation. GridsearchCV is also utilised with different ML models to help find the proper hyper-parameters that increase model accuracy. Pixel values are also collected from images without .fatten() function, keeping the data in the desired shape (244,244, 3), which is later fed into a CNN model. X_validation is created only for training the CNN model, and the data here is utilised in different epochs to aid in training a model with reasonable accuracy.

4.3 Landmark Detection Features

68 Landmark detection points are extracted from the images and used directly to infer emotion, or further processing is done to calculate the distance between points, which can also be used to classify emotions. Here the researcher utilizes the library Mlxtend. Landmark features are used normalized and not normalized to achieve and compare results. Only traditional ML and MLP were created with Landmark Detection points. Here the utilization of stratified X_train and X_test is also tested with the different transformations applied, such as PCA, LDA and data normalization. GridsearchCV is performed to aid in model fine-tuning. Figure 3 displays facial landmarks mapped and overlapped in the subject's face.

4.4 Convolutional Neural Network Design

For RGB pixel feature classfication, a convolutional neural network (CNN) was created. CNN is one of the most used computer vision solutions applied to many recent papers over the decades. The main goal of creating a CNN for this study is to compare its performance to the traditional machine learning algorithms. When using landmark detection, most libraries extract features in a way that makes them unusable in a CNN, and different CNN-friendly approaches have been handcrafted to allow CNNs to automatically import landmarks such as the paper by He et al.(2017).

4.5 Design Specification Conclusion

This concludes the Design Specification section of this research paper. Features, datasets, and literature reviews were essential in project architecture and design decisions. These decisions will aid in answering the research question. Figure 4 displays the decisions from data reading to assigning it to train and test splits.

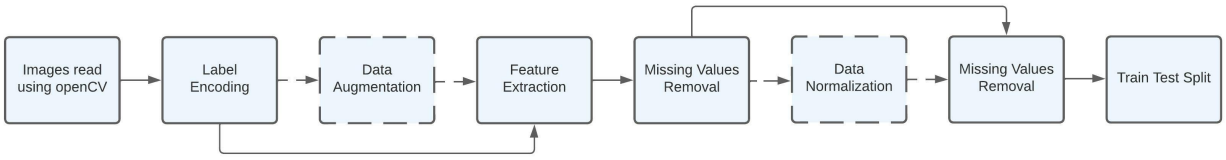


Figure 4: Preprocessing and Transformation Design. Dotted lines represent steps that were applied to some iterations but not to all

(a) JAFFE Class Balance

(b) SFEW Class Balance

Figure 5: Classes distribution after train data is split

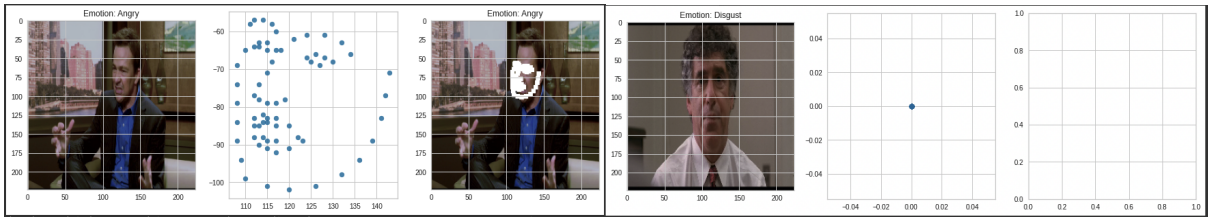
5 Implementation

This section provides details of the project's implementation and focuses mainly on the details after the pre-processing and data transformation steps are completed. Implementation discusses some of the difficulties, the multiple models developed and outlines outputs produced and the final stages of the implementation.

5.1 Classes Balance and Issues With Faces Not Being Detected

JAFFE image classes are balanced but have a small number of samples. This can be visualised in Figure 5. It is possible to train ML models, but deep learning models require a much larger number of examples to train a model properly. For the CNN model, data augmentation is used by utilising Keras sequential layers that can be used to process images. Depending on the feature selected, be that pixel values, Facial Landmarks or Facial landmark distances, the number of features varies both in the shape if .flatten() is used or not. In contrast, SFEW dataset classes, as displayed in figure 5, are imbalanced, which impacts the training and testing of the models. Data imbalance was initially addressed by using data augmentation and stratification, but feature extraction difficulties made this step redundant.

Mlxtend had difficulties finding faces in images, as shown in figure 6, an issue that



(a) SFEW face detected (b) SFEW No face detected

Figure 6: Mxtend unable to find face on all image examples

impacted SFEW examples. Due to the nature of the dataset, different brightness, occlusion and variety of face angles and artefacts, landmark detection was only able to return over six thousand examples out of 35000 augmented images, leaving over 28000 images without facial detection. To test if the data augmentation caused this issue, the same feature extraction was run on the dataset without augmentation and out of 1394 images, only 270 faces were detected.

5.2 Machine Learning Implementation

Once the data is split into Train and Test, it is ready to be either transformed by dimensionality reduction or to be used to train and test an algorithm. The algorithms were run with different features: Normalised variables, PCA, and LDA were used to compare results and aim to achieve the highest possible accuracy. Data was also shuffled to avoid models being trained in a specific order of classes which could impact the models' performance.

Some of the ML algorithms went through gridsearchCV to fine-tune the hyperparameters.

5.2.1 K-nearest neighbors

KNN was one of the models highlighted in the literature(Dino and Abdulrazzaq; 2019; Thannoon et al.; 2018) which weighted in the decision to have this algorithm selected to be part of this study and to have hyperparameters optimised by gridsearchCV. The grid parameters for number of neighbours are (3,5,11,15,19,23,25,33), Weights('uniform', 'distance') and Metrics ('euclidean', 'manhattan').

5.2.2 Random Forest

Random forest was also selected and hyperparameters were adjusted by utilizing a grid search. The different parameters are Bootstrap set to 'True' or 'false', max depth (10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None), max number of features (auto', 'sqrt'), minimal samples leaf (1, 2, 4), minimal samples split (2, 5, 10) and number of estimators (200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000).

5.2.3 Support Vector Machine

SVM is also highlighted by two papers in the literature review(Dino and Abdulrazzaq; 2019; Thannoon et al.; 2018) which also contributed to the decision of having SVM to be

compared in this research. A function was used to fit different Kernel types ('Polynomial', 'RBF', 'Sigmoid','Linear') and best Kernels results (Poly, Linear) were then used in a grid search with further parameters C of 0.001, 0.1, 1, 10, 100 and Gamma of 1, 0.1, 0.01, 0.001.

Models were run with their baseline configuration and with grid search utilizing parameter grid as per the explanation above. Both accuracy and F1 scores will be utilized to compare different models.

5.3 Convolutional Neural Network Implementation

The architecture of the CNN created for this study has three Convolutional layers. The first layer is the input layer, a Keras Conv2d, with 16 filters, 3x3 kernel size and activation 'relu'. It expects an input shape of 244 by 244 pixels and three channels (RGB). It passes through MaxPooling, which halves the output shape. The second Conv2D layer has 32 filters, 3x3 kernel size, activation 'relu', MaxPool2d, followed by batch normalization. The third Conv2D layer has 16 filters, 3x3 kernel size, activation 'relu', MaxPool2d, 0.2 dropout, followed by batch normalization, which helps control overfitting. Here the output shape is (None, 26, 26, 16). The layers are flattened, and a fully connected dense layer with 512 filters and activation 'relu' and the final dense layer with activation 'softmax' classifies the output shape into seven classes. Figure 7 shows the CNN model plot and compiled model.

6 Evaluation

The evaluation section discusses the performance of the applied algorithms. This study used a confusion matrix to compare different models and results. As data was divided into training, test and for the CNN validation splits, the test data is used to evaluate the models' performance. The Confusion Matrix graphs contain information on Precision, Recall, F1-score, Support and Accuracy, both macro and weighted average. For this study, accuracy and F1-weighted average will be called out in all experiments. F1-score weighted will be used to discuss imbalanced data as it considers the performance of correct predictions over all classes according to the number of samples in each class. Model performance is calculated based on the values present in the confusion matrix once fitting and prediction are completed. Legend: TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

$$Accuracy = (tp + tn)/(tp + fn + fp + tn)$$

$$Precision = tp/(tp + fp)$$

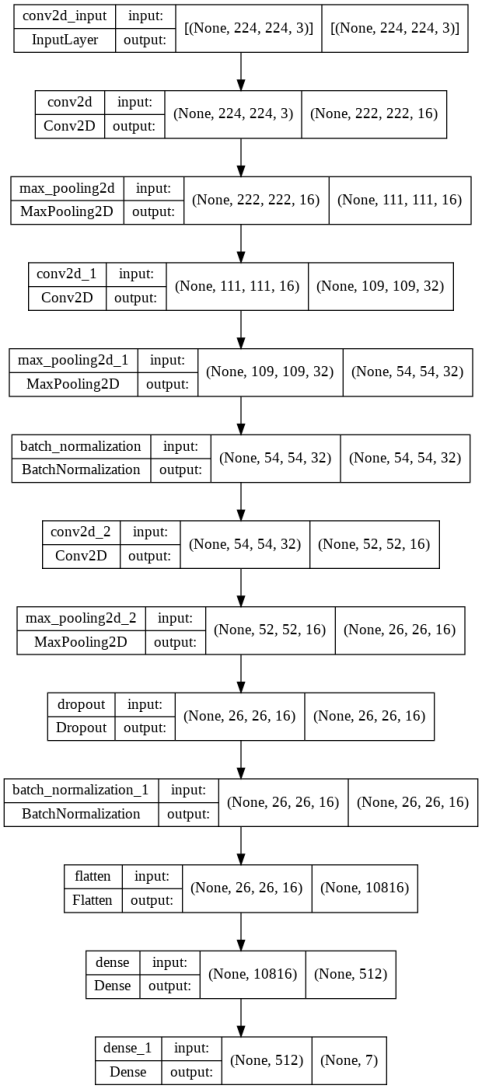
$$Recall = tp/(tp + fn)$$

$$F1 - Score = 2x(precision * recall)/(precision + recall)$$

6.1 Experiment 1 / Pixel Values

Pixel values were extracted from the images and flattened. After that, three different approaches were experimented with: normalised pixel values, LDA values and a PCA to reduce dimensionality and see if the results were further improved.

CNN Model



(a) CNN model plot

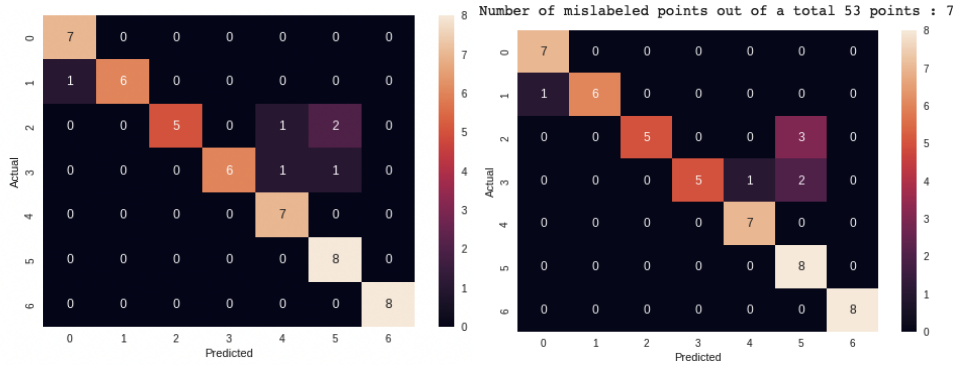
Layer (type)	Output Shape	Param #
conv2d_45 (Conv2D)	(None, 222, 222, 16)	448
max_pooling2d_45 (MaxPooling2D)	(None, 111, 111, 16)	0
conv2d_46 (Conv2D)	(None, 109, 109, 32)	4640
max_pooling2d_46 (MaxPooling2D)	(None, 54, 54, 32)	0
batch_normalization_22 (BatchNormalization)	(None, 54, 54, 32)	128
conv2d_47 (Conv2D)	(None, 52, 52, 16)	4624
max_pooling2d_47 (MaxPooling2D)	(None, 26, 26, 16)	0
dropout_18 (Dropout)	(None, 26, 26, 16)	0
batch_normalization_23 (BatchNormalization)	(None, 26, 26, 16)	64
flatten_15 (Flatten)	(None, 10816)	0
dense_34 (Dense)	(None, 512)	5538304
dense_35 (Dense)	(None, 7)	3591

Total params: 5,551,799
Trainable params: 5,551,703
Non-trainable params: 96

(b) Compiled CNN Model

Figure 7: CNN Compiled model layers, output shape and parameters

JAFFE Best Results Pixel Values



	precision	recall	f1-score	support
0	0.88	1.00	0.93	7
1	1.00	0.86	0.92	7
2	1.00	0.62	0.77	8
3	1.00	0.75	0.86	8
4	0.78	1.00	0.88	7
5	0.73	1.00	0.84	8
6	1.00	1.00	1.00	8
accuracy			0.89	53
macro avg	0.91	0.89	0.89	53
weighted avg	0.91	0.89	0.88	53

(a) Ensemble Model 89%

	precision	recall	f1-score	support
0	0.88	1.00	0.93	7
1	1.00	0.86	0.92	7
2	1.00	0.62	0.77	8
3	1.00	0.62	0.77	8
4	0.88	1.00	0.93	7
5	0.62	1.00	0.76	8
6	1.00	1.00	1.00	8
accuracy			0.87	53
macro avg	0.91	0.87	0.87	53
weighted avg	0.91	0.87	0.87	53

(b) Extra Trees Classifier 87%

Figure 8: Best results for JAFFE pixel feature

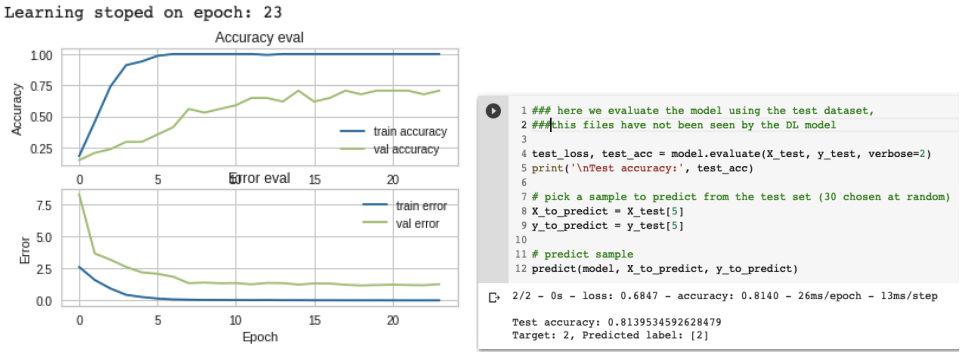
6.1.1 JAFFE Dataset

On the JAFFE dataset, support shows balance amongst classes, so the overall average will be used. KNN results were fine-tuned by performing a grid search, and an accuracy of 68% was observed with the normalised data and 75% accuracy with LDA baseline results. Random Forest Classifier normalised results had an accuracy of 79% (The fine-tuning of hyperparameters added 1% to the weighted average F1-score) and LDA results of 85% accuracy. Extra Trees Classifier achieved accuracy of 87% for normalised data and 79% for LDA. SVM achieved overall accuracy of 85% for normalised data and 77% for LDA. PCA results achieved much smaller accuracy results in a Extra Trees Classifier test with 34% accuracy. The full confusion matrix of the best-achieved results are displayed in figure 8. An ensemble model was created with models with similar F1-scores to try to increase the maximum achieved accuracy. Utilising hard voting, an ensemble constructed using Random Forest, Extra Trees Classifier and SVM with grid search configuration was combined, and the results increased overall accuracy from 87% with the Extra Trees classifier to 89% with the ensemble model. Figure 8 displays the best results for the JAFFE dataset.

6.1.2 JAFFE CNN Results

The constructed CNN network was compiled with Adam optimizer, a learning rate of 0.0001 and sparse categorical crossentropy loss. An early stop was also created to stop training in case of overfitting, with patience set to 5. After the model completed training, it was used to predict the Test split of the data that it had not seen before. The results show an accuracy of 81%. There's evidence of overfitting based on the produced accuracy and difference between training, validation and testing datasets. Particularly for posed datasets overfitting is a common problem due to the near perfect conditions in which

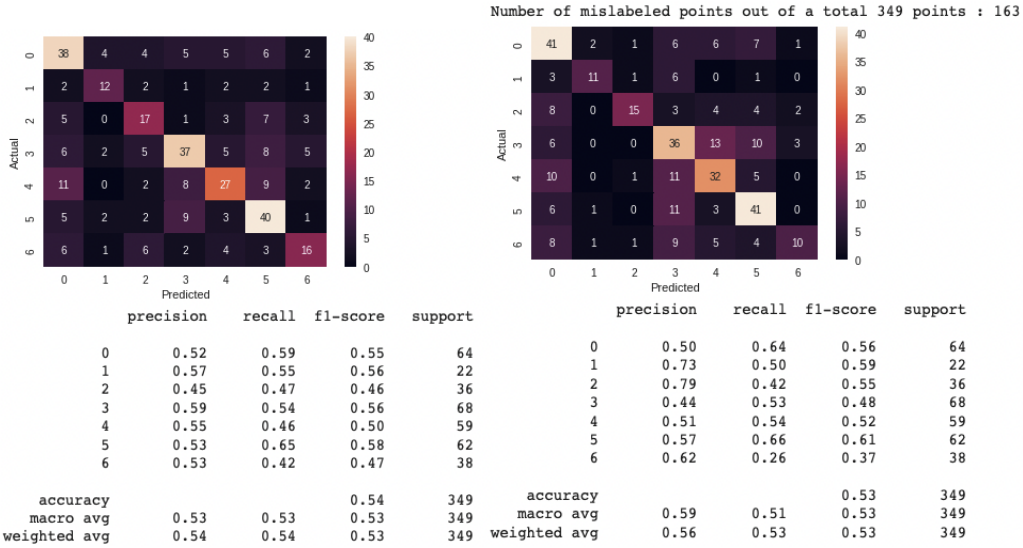
CNN Results



(a) History Plotting (b) Test Accuracy

Figure 9: CNN layers, output shape, parameters and Resulting Accuracy

SFEW Best Results Pixel Values



	precision	recall	f1-score	support
0	0.52	0.59	0.55	64
1	0.57	0.55	0.56	22
2	0.45	0.47	0.46	36
3	0.59	0.54	0.56	68
4	0.55	0.46	0.50	59
5	0.53	0.65	0.58	62
6	0.53	0.42	0.47	38
accuracy			0.54	349
macro avg	0.53	0.53	0.53	349
weighted avg	0.54	0.54	0.53	349

(a) KNN

	precision	recall	f1-score	support
0	0.50	0.64	0.56	64
1	0.73	0.50	0.59	22
2	0.79	0.42	0.55	36
3	0.44	0.53	0.48	68
4	0.51	0.54	0.52	59
5	0.57	0.66	0.61	62
6	0.62	0.26	0.37	38
accuracy			0.53	349
macro avg	0.59	0.51	0.53	349
weighted avg	0.56	0.53	0.53	349

(b) Extra Trees

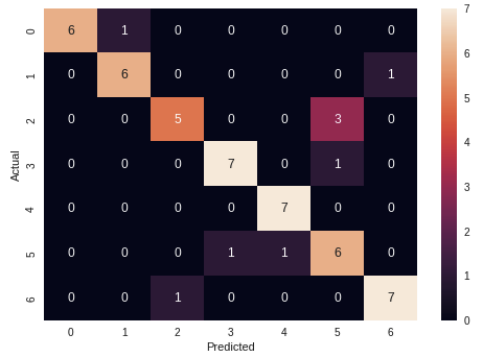
Figure 10: Best Results for SFEW Dataset and Pixel Feature

the datasets are produced which impact model ability to generalise information. Model accuracy can be different than real life accuracy. Figure 9 shows both the training history and results achieved with the model test.

6.1.3 SFEW Dataset

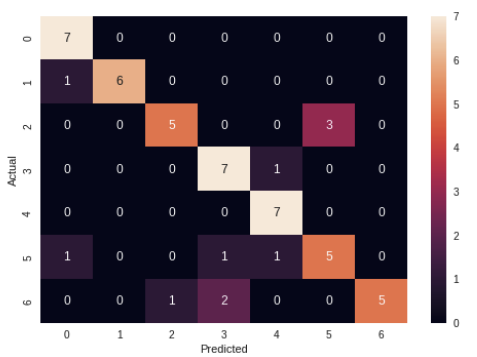
The SFEW dataset KNN results had a max accuracy of 54% and a weighted F1-score of 53%. With LDA transformation KNN results achieved both accuracy and an F1-score of 38%. Extra trees classifier achieved 53% F1-score and 35% with LDA data transformation. Random Forest Achieved 47% F1-score and 48% overall accuracy. Other algorithms achieved lower scores. Figure 10 shows the best SFEW results, KNN and Extra Trees.

JAFFE Best Results Landmark Points



	precision	recall	f1-score	support
0	1.00	0.86	0.92	7
1	0.86	0.86	0.86	7
2	0.83	0.62	0.71	8
3	0.88	0.88	0.88	8
4	0.88	1.00	0.93	7
5	0.60	0.75	0.67	8
6	0.88	0.88	0.88	8
accuracy			0.83	53
macro avg	0.85	0.83	0.83	53
weighted avg	0.84	0.83	0.83	53

(a) SVM 83%



	precision	recall	f1-score	support
0	0.78	1.00	0.88	7
1	1.00	0.86	0.92	7
2	0.83	0.62	0.71	8
3	0.70	0.88	0.78	8
4	0.78	1.00	0.88	7
5	0.62	0.62	0.62	8
6	1.00	0.62	0.77	8
accuracy			0.79	53
macro avg	0.82	0.80	0.79	53
weighted avg	0.81	0.79	0.79	53

(b) Random Forest Classifier 79%

Figure 11: Best results for JAFFE landmark points

6.2 Experiment 2 / Landmark Detection Points

6.2.1 JAFFE Dataset

KNN results accuracy peaked at 53%. KNN results of 45% accuracy were observed with LDA baseline results. Random Forest Classifier Results had an accuracy of 79% and LDA results of 36% accuracy. Extra Trees Classifier achieved an accuracy of 75% and an accuracy of 40%. SVM achieved an overall accuracy of 83% and 43% for LDA. PCA results achieved accuracy results in an SVM Classifier test with on 32% accuracy. Figure 11 shows the confusion matrix for the models with the best accuracy.

6.2.2 SFEW Dataset

SFEW landmark points achieved lower results compared to the results achieved by pixel values. The highest accuracy was generated by utilizing Random Forest with 45% F1-score followed by 40% in extra trees. Both LDA techniques and PCA resulted in lower accuracy.

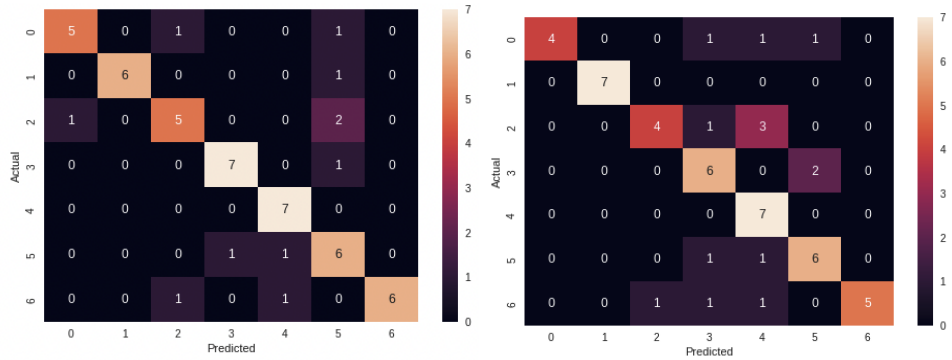
6.3 Experiment 3 / Landmark Distances

After the landmark detection points were collected the euclidean distance between points was calculated and used as feature.

6.3.1 JAFFE Dataset

JAFFE dataset KNN results Accuracy peaked at 53% and LDA results of 55% accuracy. Random Forest Classifier Results achieved an accuracy of 64%, LDA results of 57% accuracy. Extra Trees Classifier achieved accuracy and weighted f1-score of 74% and

JAFFE Best Results Landmark Distances



	precision	recall	f1-score	support
0	0.83	0.71	0.77	7
1	1.00	0.86	0.92	7
2	0.71	0.62	0.67	8
3	0.88	0.88	0.88	8
4	0.78	1.00	0.88	7
5	0.55	0.75	0.63	8
6	1.00	0.75	0.86	8
accuracy			0.79	53
macro avg	0.82	0.80	0.80	53
weighted avg	0.82	0.79	0.80	53

(a) SVM 79%

	precision	recall	f1-score	support
0	1.00	0.57	0.73	7
1	1.00	1.00	1.00	7
2	0.80	0.50	0.62	8
3	0.60	0.75	0.67	8
4	0.54	1.00	0.70	7
5	0.67	0.75	0.71	8
6	1.00	0.62	0.77	8
accuracy			0.74	53
macro avg	0.80	0.74	0.74	53
weighted avg	0.80	0.74	0.74	53

(b) Extra Trees Classifier 74%

Figure 12: Best results for JAFFE landmark distances

36% weighted F1-score for LDA with an accuracy of 58% overall. SVM achieved an overall accuracy of 79% on accuracy and 57% for LDA. PCA achieved accuracy results in an SVM Classifier test with 28% accuracy. Figure 12 displays the confusion matrix for the models with the best accuracy.

6.3.2 SFEW Dataset

SFEW results for landmark distances were lower than the pixel value features, peaking at 35% F1-score with SVM. Both PCA and LDA techniques applied to the dataset resulted in lower scores.

6.4 Discussion

Table 2 provides an overview of the ten best-performing models trained with JAFFE by accuracy; it includes the features used and the resulting accuracy and weighted F1-score. The results achieved using the JAFFE dataset are satisfactory, with the highest accuracy achieved by the ensemble model with 89% accuracy on the normalised pixel value feature. PCA tests resulted in considerably smaller accuracy results, while LDA increased accuracy for KNN pixel results compared to pixel feature data that was only normalised but not LDA transformed. Looking at the 10 best performing models' accuracy, the pixel value feature appears on the list 7 out of 10 times with an accuracy range between 77% and 89%, landmark points appear twice with an accuracy of 79% and 83% and landmark distances once with an accuracy of 79%.

Table 2: JAFFE 10 Best Models per F1-Score

Algorithm	Feature	JAFFE	
		Accuracy	**F1-Score**
Ensemble	Pixel Value	89%	88%
Extra Trees	Pixel Value	87%	87%
Random Forest (LDA)	Pixel Value	85%	85%
SVM	Pixel Value	85%	85%
SVM	Landmark Points	83%	83%
CNN	Pixel Value	81%	-
Random Forest	Pixel Value	79%	79%
Extra Trees (LDA)	Pixel Value	79%	79%
Random Forest	Landmark Points	79%	79%
SVM	Landmark Distances	79%	80%

The CNN constructed achieved test accuracy of 81% while both ensemble and machine learning models such as Extra Trees and SVM achieved higher accuracy with 89% and 85%. Results from traditional machine learning are comparable to deep learning results, which are displayed in table 1. Overall, pixel value features the best accuracy outperformed landmark points' best accuracy scores by 6 points.

Table 3 provides an overview of the ten best performing models trained with SFEW by F1-score; it includes the features used and resulting accuracy and weighted F1-score. SFEW dataset achieved the highest accuracy, and F1-Score with Pixel values features achieved 53% with Extra Trees classifier, Followed by KNN with an F1-score of 53% and 54% overall accuracy. On Landmark points, the best result was with Random Forest achieving an F1-score of 45%. The calculation of euclidean distances between landmark points resulted in a much smaller range of results, between F1-scores of 35% and 30% for normalised data and more minor results for both LDA and PCA transformed data.

Table 3: SFEW 10 Best Models per F1-Score

Algorithm	**Feature**	**SFEW**	
		Accuracy	**F1-Score**
KNN	Pixel Value	54%	53%
Extra Trees	Pixel Value	53%	53%
Random Forest	Pixel Value	48%	47%
Random Forest	Landmark Points	47%	45%
SVM	Pixel Value	44%	44%
Extra Trees	Landmark Points	43%	40%

Table 3: SFEW 10 Best Models per F1-Score

Algorithm	**Feature**	**SFEW**	
		Accuracy	**F1-Score**
KNN	Landmark Points	40%	39%
KNN (LDA)	Pixel Value	38%	38%
Random Forest (LDA)	Pixel Value	37%	37%
SVM (LDA)	Pixel Value	37%	37%

For JAFFE, the range between features is as follows. Pixel Value feature 68% to 89%. Landmark points with a range between 83% and 36% with normalised only data in the top four results (range between 53% 83%) and LDA results after (45% and 36%). Landmark points distances between 79% and 53%. PCA results achieved the worse accuracy between 34% and 28%. For the SFEW dataset, LDA transformed data had lower accuracy than just normalised data which can be visualised in all features as per Table 4.

Table 4: Ranges between Features

Feature	**JAFFE**	**SFEW**
	Results Accuracy Range	
Pixel Value	68% - **89%**	Normalised 44% - **53%**, LDA 35% - 38%
68 Landmark Points	Normalised 53% - **83%**, LDA 36% - 45%	Normalised 34% - **45%**, LDA 18% - 24%
Landmark Distances	53% - **79%**	Normalised 30% - **35%**, LDA 13% - 19%

Landmark detection in this study did not increase the overall accuracy of models, and even in the best performing models and dataset, the usage of Landmark detection was 6 points short of the results achieved using pixel values, 10 points difference was observed between low level feature pixel values and landmark distances in JAFFE. The decision to use posed or spontaneous datasets can highly impact the results achieved when creating solutions for FER. The JAFFE dataset has positive and negative implications depending on the angle a researcher might look at it. It was created with near perfect conditions, clear background, no distractions from the subject's face, good lighting, good posing and no occlusions. The same points that make this dataset near perfect for training make it hard to translate to real-life scenarios. SFEW examples are more diverse, with examples closer to real-life scenarios, but this also increases the difficulty in obtaining higher accuracy.

7 Conclusion and Future Work

This study aimed to build different machine learning models that can effectively classify human emotions. Answering the proposed research questions on how traditional machine learning techniques compare to deep learning solutions in accuracy, verify if the usage of landmark detection points increases model accuracy and test which of the selected test models outperform others when classifying FER. Results demonstrate that traditional machine learning techniques such as SVM, Extra Trees classifier and ML ensemble can achieve results comparable to some of the deep learning models preferred by researchers and some of the results displayed in the related works section. Landmark detection, in this study, did not outperform models created utilising pixel values extracted from images and the best-produced model that outperformed others was a model ensemble which achieved 89% accuracy utilising pixel values as its features. This research showcases the usage of traditional ML to solve FER problems with fewer resource needs, are quick to implement and does not require a large number of examples to produce reasonably good accuracy. It displays the impact that the type of dataset chosen by research has on the produced results and how posed datasets' performance is much higher to spontaneous datasets. Single culture/etnicity datasets such as JAFFE have less variation than a multicultural dataset such as SFEW. Datasets are created different and both have their value particularly when looking from the point that basic emotions mentions that emotions are displayed in the same way over different cultures. SFEW has over 60 subjects of various ethnic groups and a model must be able to generalise well to accommodate different features, beards and artefacts. While the model created using posed datasets might have more generalisation issues and overfitting, businesses can try to recreate the favourable conditions in which the posed datasets are created to enhance results in real life, such as adjusting brightness and cropping background after the input image is collected before emotion classification takes place.

This work can be improved for future works by focusing on data transformation and augmentation, trying to recreate the conditions present in a posed dataset in "in-the-wild" datasets. Using different points or calculations for point distances might also be helpful as some of the results reported in different papers achieved higher accuracy than the ones in this study. Researchers should also consider utilising different landmark detection libraries that might be more successful extracting facial points information from diverse images. Further research is also needed to understand better emotions and how to label them successfully.

8 Acknowledgments

I want to thank NCI and supervisor, Prof. Jorge Basilio, for his guidance and help in requesting datasets, discussing the research problem and providing direction during the entire process of completing this research.

References

Alqahtani, F., Katsigiannis, S. and Ramzan, N. (2019). Ecg-based affective computing for difficulty level prediction in intelligent tutoring systems, *2019 UK/ China Emerging Technologies (UCET)*, pp. 1–4.

Aranha, R. V., Corrêa, C. G. and Nunes, F. L. S. (2021). Adapting software with affective computing: A systematic review, *IEEE Transactions on Affective Computing* **12**(4): 883–899.

Barman, A. and Dutta, P. (2017). Facial expression recognition using shape signature feature, *2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, pp. 174–179.

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E. and D'Mello, S. K. (2021). Integrating psychometrics and computing perspectives on bias and fairness in affective computing: A case study of automated video interviews, *IEEE Signal Processing Magazine* **38**(6): 84–95.

Canabal, M. F., Miranda, J. A., Lanza-Gutiérrez, J. M., Pérez Garcilópez, A. I. and López-Ongil, C. (2020). Electrodermal activity smart sensor integration in a wearable affective computing system, *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*, pp. 1–6.

Cao, H. and Qi, C. (2021). Facial expression study based on 3d facial emotion recognition, *2021 20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS)*, pp. 375–381.

Cho, H., Kang, W. K., Park, Y.-S., Chae, S. G. and Kim, S.-j. (2022). Multi-label facial emotion recognition using korean drama video clips, *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 215–221.

Chouhayebi, H., Riffi, J., Mahraz, M. A., Yahyaouy, A. and Tairi, H. (2021). Facial expression recognition using machine learning, *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pp. 1–6.

Comas, J., Aspandi, D. and Binefa, X. (2020). End-to-end facial and physiological model for affective computing and applications, *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 93–100.

Dhall, A., Goecke, R., Lucey, S. and Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies, *IEEE multimedia* **19**(03): 34–41.

Dino, H. I. and Abdulrazzaq, M. B. (2019). Facial expression classification based on svm, knn and mlp classifiers, *2019 International Conference on Advanced Science and Engineering (ICOASE)*, pp. 70–75.

Haghpanah, M. A., Saeedizade, E., Masouleh, M. T. and Kalhor, A. (2022). Real-time facial expression recognition using facial landmarks and neural networks, *2022 International Conference on Machine Vision and Image Processing (MVIP)*, IEEE, pp. 1–7.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, Z., Kan, M., Zhang, J., Chen, X. and Shan, S. (2017). A fully end-to-end cascaded cnn for facial landmark detection, *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pp. 200–207.

King, D. E. (2009). Dlib-ml: A machine learning toolkit, *J. Mach. Learn. Res.* **10**: 1755–1758.

Lyons, M. J. (2021). " excavating ai" re-excavated: Debunking a fallacious account of the jaffe dataset, *arXiv preprint arXiv:2107.13998* .

Lyons, M. J., Kamachi, M. and Gyoba, J. (2020). Coding facial expressions with gabor wavelets (ivc special issue), *arXiv preprint arXiv:2009.05938* .

Masson, A., Cazenave, G., Trombini, J. and Batt, M. (2020). The current challenges of automatic recognition of facial expressions: A systematic review, *AI Communications* **33**(3-6): 113–138.

Maurya, A. and Sharma, V. (2022). Facial emotion recognition using keras and cnn, *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pp. 2539–2543.

Ong, D. C. (2021). An ethical framework for guiding the development of affectively-aware artificial intelligence, *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, pp. 1–8.

Qiu, Y. and Wan, Y. (2019). Facial expression recognition based on landmarks, IEEE, pp. 1356–1360.

Shen, X., Hu, X., Liu, S., Song, S. and Zhang, D. (2020). Exploring eeg microstates for affective computing: decoding valence and arousal experiences during video watching¡sup¿*¡/sup¿, *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 841–846.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* .

Stanciu, L. and Albu, A. (2019). Analysis on emotion detection and recognition methods using facial microexpressions. a review, *2019 E-Health and Bioengineering Conference (EHB)*, pp. 1–4.

Stark, L. and Hoey, J. (2021). The ethics of emotion in artificial intelligence systems, *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 782–793.

Takalkar, M. A. and Xu, M. (2017). Image based facial micro-expression recognition using deep learning on small datasets, IEEE, pp. 1–7.

Thannoon, H. H., Ali, W. H. and Hashim, I. A. (2018). Detection of deception using facial expressions based on different classification algorithms, *2018 Third Scientific Conference of Electrical Engineering (SCEE)*, pp. 51–56.

Wang, C., Peng, M., Bi, T. and Chen, T. (2020). Micro-attention for micro-expression recognition, *Neurocomputing* **410**: 354–362.

Wang, S., Fang, Z., Zhang, S. and Dong, H. (2021). Research on affective computing based on graph sentiment dictionary, *2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST)*, pp. 793–797.

Wessler, J., Schneeberger, T., Hilpert, B., Alles, A. and Gebhard, P. (2021). Empirical research in affective computing: An analysis of research practices and recommendations, IEEE, pp. 1–8.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining, *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, Vol. 1, Manchester, pp. 29–39.

Wu, Y. and Ji, Q. (2019). Facial landmark detection: A literature survey, *International Journal of Computer Vision* **127**(2): 115–142.

Özseven, T. and Düğenci, M. (2017). Face recognition by distance and slope between facial landmarks, *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–4.