National College of Ireland

BSc. (Hons.) Data Science

Final Year Project

2022

Lauren Kelly

19362533

X19362533@student.ncirl.ie

Supervisor: Dr. Arghir Moldovan

"A Comparative Analysis of Emotion Detection Accuracy in AWS Rekognition and Luxand FaceSDK: Balancing Performance and Cost"

Contents

# Acknowledgements

First and foremost, I'd like to express my gratitude to my supervisor, Dr. Arghir Moldovan, for his ongoing guidance and support over the past few months. His feedback and advice were extremely helpful in completing this final year project and are greatly appreciated.

I would also like to thank the Computing Support Team at NCI for granting me access to systems and resources that were extremely useful in completing this work.

Thank you especially to my mother and father, Elaine McConnell and Gavin Kelly, without whom this would not have been possible and whose love and encouragement guided me throughout this entire BSc program.

Finally, I'd like to thank the School of Computing at the National College of Ireland Dublin for their help and support over the last four years.

Executive Summary

In this final year project, a comparison is made with the performance of two cloud-based vision applications, Amazon Web Services (AWS) Rekognition and Luxand FaceSDK, in detecting emotions from a dataset of labelled images. The dataset was compiled by amalgamating three publicly available datasets, providing a diverse and representative sample for testing the emotion detection capabilities of these APIs.

The study aimed to analyse the accuracy of these two platforms, which possess different cost structures for the users, to help stakeholders, make informed decisions when choosing a facial recognition API for their applications.

To conduct this comparison, custom Python code was developed to iterate over the images, call the respective cloud services, and save the emotion detection results to JSON files. The accuracy of the APIs was determined by comparing their outputs to the labels provided within the original dataset.

Face recognition software must be accurate because it has wide-ranging effects on a variety of industries, including security, marketing, and entertainment. To accurately identify people and grant or deny access, security systems, for instance, rely on facial recognition software. Marketing campaigns can be more focused and successful if they consider the emotional reactions of consumers to advertisements. Emotion detection can be used in interactive experiences like virtual reality or gaming in the entertainment industry to increase user engagement.

In conclusion, this project provides a comprehensive comparison of the AWS Rekognition and Luxand FaceSDK platforms in terms of emotion detection accuracy. The results will enable potential users to make informed decisions based on their specific requirements and budget constraints while emphasizing the significance of accuracy in facial recognition applications across various domains.

1.0 Introduction

1.1. Background

Facial expression recognition is an area of research that has gained widespread interest due to its potential applications in various fields, including computer vision, machine learning, human-computer interaction, neuroscience, psychology, cognition system, and sports.

Thanks to advancements in computer vision techniques, facial expression recognition has become an area of active research. Tech giants like Amazon, Google, and Microsoft Corporation are heavily investing in this technology, offering cloud-based APIs that use powerful machine learning algorithms to detect emotions from images and videos. However, the cost of accessing these services may be high and not a feasible solution for some developers and organizations.

However with the integration of facial recognition into everyday life, is it really accessible to standard people to integrate into start up applications or software? In this project, two emotion detection API's are compared; AWS Rekognition (Services, 2023) and Luxand FaceSDK (Luxand, 2023). Datasets full of pre-labelled images were downloaded from publically available locations and combined to test the functionalities of these API's.

1.2. Aims

This research aims to provide a detailed comparison of the accuracy of these two platforms when applied to a dataset composed of labelled images sourced from three publicly available datasets. This dataset was meticulously curated to ensure a balanced and representative sample for the evaluation of the APIs' emotion detection capabilities. The project's primary goal is to assist stakeholders in making informed decisions when choosing a facial recognition API for their applications, while considering the trade-offs between performance and cost.

AWS Rekognition is a cloud-based service provided by Amazon Web Services (AWS) that offers powerful image and video analysis capabilities using deep learning algorithms. It enables developers to easily integrate computer vision functionalities into their applications without requiring deep expertise in machine learning or computer vision. With AWS Rekognition, one can perform various tasks such as object and scene detection, facial analysis, facial recognition, text detection, and moderation of explicit or inappropriate content. The service provides APIs that allow you to integrate these capabilities into your applications, making it easier to automate image and video analysis workflows. For this project, the focus will be on it's emotion detection capabilities. The pricing for AWS Rekognition varies based on the specific features and the amount of usage. The pricing is typically divided into two components: 1) API requests and 2) storage. The API request pricing includes charges for various operations like image analysis, facial recognition, and text detection. The storage pricing applies if you choose to store images or videos in the Rekognition service for long-term usage.

For those seeking more cost-effective options, the Luxand facial expression recognition software development kit (SDK) is a viable alternative. Although it is a commercial product, the Luxand SDK may be more cost-effective than cloud-based APIs for some developers and organizations. The Luxand SDK provides powerful facial expression recognition capabilities and other computer vision tools that can be licensed for commercial use. In addition, there are open-source facial expression recognition libraries like OpenFace and Affectiva Emotion SDK that can be used without incurring high costs for non-commercial and research purposes.

1.3. Technology

The project employs custom Python code to automate the process of iterating through the dataset, invoking the respective cloud services, and recording the emotion detection results in CSV and JSON files. These outputs are then compared to the original labels provided within the dataset to determine the accuracy of each platform.

Python code was also written to divide the datasets into sub-folders for analysis, within these subfolders each complete image output is written to its own JSON file and a line is added to the CSV for each image that shows the confidence levels for each emotion. This code was then adapted to create subsets for the analysis in Luxand.

I utilised Luxand FaceSDK's free trial to use the API and I accessed AWS Rekognition through NCI's cloud services. With the Luxand free trial, I was limited to 50,000 transactions – which encouraged the creation of the subsets which will be discussed later in the paper. With the cloud access for AWS Rekognition, I was not limited to any number of transactions. Therefore, I processed all the images through AWS Rekognition but only processed a subset through Luxand FaceSDK.

There were other things revealed in the Rekognition output, such as gender detection, detection of accessories etc. which is why it was fundamental to include the full output as a JSON, to potentially be used in a future study.

1.4. Structure

We will go over the data collection and analysis techniques used in this report, present the comparison's results, and provide some insight into how these findings might affect different industries that use emotion recognition. We will also emphasize the critical importance of accuracy in facial recognition systems by citing instances from the security, marketing, and entertainment industries, where accurate emotion recognition can have a big impact on the usability and success of applications. This will take the following structure:

1. State of the Art
   A summary of the most recent developments and methods for emotion detection using computer vision and deep learning techniques. It discusses the importance of emotion recognition in a variety of fields, including affective computing and human-computer interaction. The section emphasizes the use of machine learning algorithms and facial expression analysis for automatic emotion recognition. It outlines the difficulties in emotion recognition, including the variability of facial expressions, biases in datasets, and the requirement for sizable, annotated datasets. The "State of the Art" section lays the groundwork for the subsequent analysis by outlining the state of the art in research and emphasizing the necessity of comparing various emotion detection APIs to gauge their effectiveness and potential.

2. Data
   This section contains an overview of the datasets used for the evaluation of the emotion detection API's. The Cohn-Kanade (CK), Expressions in the Wild (ExpW), and Facial Expression Recognition (FER) datasets are among the specific datasets mentioned in the section. It briefly outlines each dataset's features, including its resolution, image count, and available emotion labels. The section emphasizes how crucial it is to use a variety of representative datasets to guarantee the accuracy and generalizability of the findings. A general overview of the datasets used in the study is given in the "Data" section, laying the groundwork for the analysis and evaluation of the emotion detection APIs that follow.

3. Methodology
   This section outlines the steps taken and the overall approach to evaluate two emotion detection APIs, namely AWS Rekognition and Luxand FaceSDK with implementation of the CRISP-DM methodology. The section describes the datasets used for the analysis, including the Cohn-Kanade (CK) dataset, Expressions in the Wild (ExpW) dataset, and Facial Expression Recognition (FER) dataset. It explains the pre-processing steps, such as face detection and emotion label extraction, performed on the datasets before feeding them into the APIs. The section also discusses the metrics used to assess the performance of the APIs, including average confidence levels, binary accuracy, and failure rates. Additionally, it explains the comparison and analysis of the results obtained from the APIs across the different datasets. The "Methodology" section provides a clear framework for the experimental setup and evaluation of the emotion detection APIs.

4. Analysis
   The "Analysis" section of the paper presents a detailed evaluation and comparison of two emotion detection APIs: AWS Rekognition and Luxand FaceSDK. The section begins by discussing the average confidence levels for each emotion in the different datasets, highlighting the strengths and weaknesses of each API. It examines the accuracy and misclassification rates of the APIs in detecting dominant emotions across the datasets. The analysis includes comparisons of mean confidence levels, binary accuracy, and percentage of failure in face detection. The section also identifies notable patterns and trends, such as the higher confidence levels for certain emotions and the challenges in accurately detecting specific emotions.

5. Results
   The main findings from the analysis of the emotion detection APIs, AWS Rekognition and Luxand FaceSDK, using various datasets are presented in the section of the paper titled "Results". In terms of average confidence levels, binary accuracy, and the capacity to correctly categorize dominant emotions, it highlights how well each API performed. In terms of higher confidence levels, better binary accuracy, and lower failure rates, the results show that AWS Rekognition performs generally better than Luxand FaceSDK. The section also goes over specific findings for each dataset, like the misclassification of some emotions and performance differences between emotions. Overall, the findings show that AWS Rekognition and Luxand FaceSDK both provide reliable and robust emotion detection capabilities.

6. Conclusions
   The "Conclusions" section of the paper summarizes the key findings and outcomes of the project. It highlights that AWS Rekognition performed better than Luxand FaceSDK in terms of accuracy and confidence levels in emotion detection across the evaluated datasets. AWS Rekognition demonstrated higher confidence and accuracy in identifying dominant emotions, particularly for emotions like 'Happy' and 'Surprise', while Luxand FaceSDK had lower confidence levels and misclassifications, especially for 'Disgust'. The binary accuracy and mean confidence comparison further emphasized the reliability of AWS Rekognition in detecting emotions. The section concludes that AWS Rekognition is a robust and comprehensive emotion detection API, achieving high accuracy rates, while Luxand FaceSDK has limitations and higher failure rates.

2.0 State of the Art

Many analyses have been carried out on the facial recognition programs independently of each other. This will aid the analysis carried out in this project as different datasets were analysed in these, resulting in different conclusions. These conclusions and papers will assist in providing a broader image of the programs before use.

Anatomizing Bias in Facial Analysis (Singh, et al., 2022)

It has been demonstrated that the results of current facial analysis systems are biased against demographic subgroups. It is crucial to make sure that these systems do not discriminate against people based on their gender, identity, or skin tone because of the negative effects they have on society. Research on identifying and reducing bias in AI systems has resulted from this. This paper summarizes bias detection/estimation and mitigation algorithms for facial analysis in this paper. The main contributions include a thorough analysis of the algorithms proposed for understanding bias as well as a taxonomy and comprehensive overview of the algorithms currently in use.

Actionable Auditing: Investigating the Impact of Publically Naming Biased Performance Results of Commercial AI Products (Raji & Buolamwini, 2019)

Although algorithmic auditing has become a critical tool for exposing systematic biases in software platforms, research on how these audits affect the fairness and transparency of algorithms in commercial systems is still in its infancy. In this paper, the commercial impact of Gender Shades, the first algorithmic audit of gender and skin type performance disparities in commercial facial analysis models is examined to analyze the impact of publicly naming and disclosing performance results of biased AI systems.

Overview of the Advancements in Automatic Emotion Recognition: Comparative Performance of Commercial Algorithms (Malygina, et al., 2019)

In recent years, facial emotion recognition algorithms have advanced, and today's top commercial algorithms are sometimes better at detecting emotions like happiness than people. It is customary to assess these algorithms' performance by contrasting it with human-labelled ground truth. This article discusses tracking improvements in automatic emotion recognition systems, and here we propose an additional criterion for assessing them: the consistency of the predictions made by the algorithms. The performance of four commercial algorithms is compared in this study: Affectiva Affdex, Microsoft Cognitive Services Face module Emotion Recognition, Amazon Rekognition Face Analysis, and Neurodata Lab Emotion Recognition. Overall findings reveal that the algorithms developed by Microsoft, Neurodata Lab, and Amazon that obtained higher f1-scores and accuracy for human-labelled ground truth had greater agreement between their predictions. In order to further investigate the possibility of using automatic annotation to replace human data labelling, agreement among algorithmic predictions is a promising criterion.

Simulations Models of Face-Based Emotion Recognition (Goldman & SekharSripada, 2005)

Recent research on emotion mindreading shows that deficits in face-based recognition are correlated with deficits in the production of three emotions, fear, disgust, and anger. What kind of mindreading procedure would account for this paired deficit pattern? The compatibility of the simulation approach and the theorizing approach with the available data is assessed in this paper. It was concluded that that the simulation method provides the best analysis of the data. But how might simulation-style emotion detection use computational steps? A generate-and-test model, a reverse simulation model, a variation of the reverse simulation model that uses a "as if" loop, and an unmediated resonance model are four alternate models that are investigated here.

According to recent studies, Azure Face API is the best service for emotion recognition compared to Amazon Rekognition and Google Cloud Vision. It was found to be more accurate at recognizing emotions than the other two services. I also registered as a user for the three technologies to evaluate them myself. All three services offer a free tier. During the sign-up process, Google Cloud was the easiest as it auto-fills based on Gmail information once permission has been granted. All three services required email, phone number and payment information to verify identity. Google and Microsoft also offer an initial bonus credit at sign up.

From first use, although they are performing a similar function, these technologies are vastly different. Firstly, beginning with Google Cloud Vision. This technology supports many image formats such as JPG, PNG, GIF, BMP, WebP, ICO etc. but it only accepts images directly from Google Cloud Storage. When availing of the paid tier, a bill is accumulated based on the number of images that are processed within the service. While batch processing is supported, it is limited to 8mb per request. This technology also operates with a synchronous API. It aggregates each API in a single HTTP endpoint. The output of the system is also provided in reference to Google's Knowledge Graph.

Google's Knowledge Graph is a system that is used to enhance Google's search results with relevant information from a variety of sources. This information is presented in a box to the right of the search results and is designed to provide users with a quick and easy way to find out more about a particular topic. The Knowledge Graph draws on a wide range of sources, including Wikipedia and other online encyclopedias, dictionaries, and other sources of structured data. It is designed to provide users with a more comprehensive and useful search experience by displaying relevant information alongside the search results.

Amazon Rekognition exhibited many differences to Google's service. It supports video input, object versioning and is an asynchronous API. It is much more user-friendly and less complex to operate. It also defines one HTTP endpoint per function, in contrast to Google.

The Azure Face API supports batch processing of images. The Face API allows developers to submit a batch of images for analysis, and it will return the results for each image in the batch. This can be useful for applications that need to process large numbers of images in a single request, as it allows you to submit all the images in a single request and receive the results in a single response. It's important to note that the batch processing feature of the Face API has some limitations and restrictions. For example, the maximum number of images that can be included in a single batch is 1000, and the total size of all the images in the batch must be less than 4 MB. Additionally, the Face API has limits on the number of batch processing requests that can be made in each time. You can find more information about these limits in the Azure Face API documentation.

Darwin's Claim of Universals in Facial Expression Not Challenged (Ekman, 2014)

In this paper, they clarify that Charles Darwin never claimed that all facial expressions are universal, but rather a specific set of expressions. They highlight the extensive research that has been conducted, including studies by Ekman, Izard, Haidt, and Keltner, which have shown strong cross-cultural agreement in recognizing and labelling facial expressions from the Darwin-Tomkins set. The authors also mention studies on spontaneous facial expressions and the physiological and neurophysiological responses associated with the Darwin-Tomkins set of expressions. They argue that this body of research supports the universality of facial expressions and challenges Feldman-Barrett's claims. Additionally, they note that Darwin's emphasis on the unity of mankind, as demonstrated through shared facial expressions, remains unchallenged.

Summary of Facial Recognition Programs

	Amazon Rekognition	**Luxand Face SDK**
Free Tier Available	Yes	Yes
Free Tier Limit	5,000 images / month	50,000 transactions
Free Credit	none	N/A
Fixed Cost	$0.001 / image for first million	$19 per month
Versioning Supported	Yes	Yes
Batch Processing Supported	No	Yes
API	Asynchronous	Synchronous
Image Formats Support	• PNG • JPEG	• JPEG • PNG
Videos Supported	Yes	No
Data Output Format	JSON	JSON
Age	Yes (Range)	No
Race	No	Not supported
Gender	Yes	Yes
Emotions	9 (Angry, Calm, Confused, Disgusted, Fear, Happy, Sad, Surprised, Unknown)	6 (Anger, Disgust, Fear, Happy, Sadness, Surprise)

3.0 Data

The datasets that will be used for this evaluation are Extended Cohn-Kanade Dataset, FER-2013 and Expressions in the Wild (ExpW). These Datasets are available online to the public, free for use.

Extended Cohn-Kanade Dataset

The Extended Cohn-Kanade (CK+) dataset is a facial expression database that was created to support the development of automatic facial expression recognition systems. It consists of images and videos of people displaying a range of emotions, including happiness, sadness, anger, surprise, disgust, fear, and neutral. The dataset was created by collecting images and videos of actors and actresses posing for a series of standardized facial expressions.

The CK+ dataset is an extension of the original Cohn-Kanade (CK) dataset, which was created by the same researchers and consists of similar data. The CK+ dataset includes additional images and videos, as well as improved annotations and more detailed labeling of the facial expressions. Both datasets serve as valuable resources for training and evaluating facial expression recognition models.

FER-2013 Dataset

FER-2013 is a facial expression recognition dataset that was created to support the development of automatic facial expression recognition systems. It consists of over 35,000 images of people displaying seven basic emotions (happiness, sadness, anger, surprise, fear, disgust, and neutral) as well as a "neutral" expression. The images were collected from a variety of sources, including publicly available web images and studio-generated images, and were annotated with labels indicating the presence or absence of each emotion.

One notable feature of the FER-2013 dataset is that it is relatively small compared to some other facial expression datasets, which can make it easier to use for developing and testing algorithms. However, it is also relatively limited in terms of the diversity of the images and the range of emotions represented, which may limit its usefulness for certain tasks and research goals.

ExpW (Expressions in the Wild)

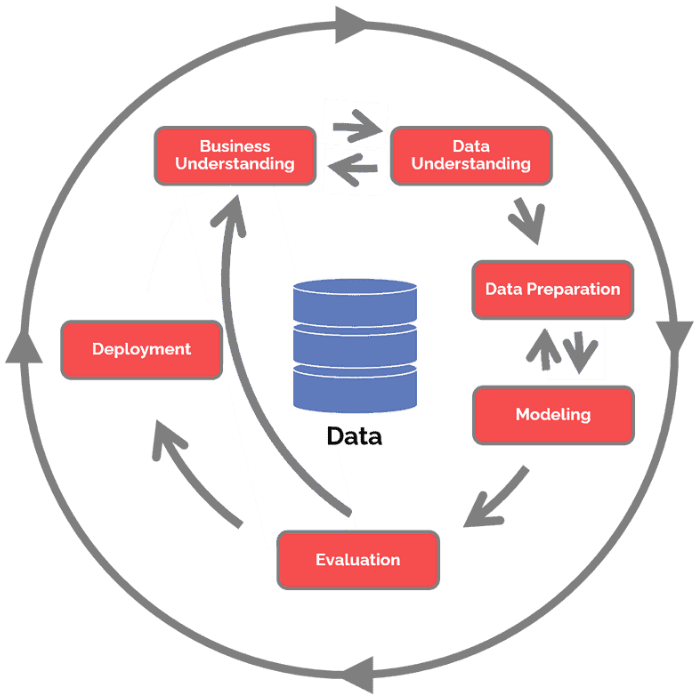
The Expression in-the-Wild (ExpW) dataset is for facial expression recognition and contains 91,793 faces manually labeled with expressions. Each of the face images is annotated as one of the seven basic expression categories: “angry”, “disgust”, “fear”, “happy”, “sad”, “surprise”, or “neutral”. (al., 2015)

Summary Table of Datasets

	Extended Cohn-Kanade Dataset	FER-2013	ExpW (Expressions in the Wild)
Content	Images and videos	Images	Images
Size	5,876 labelled images of 123 individuals	Over 35,000 images	Over 91,000 images
Emotions	6 (happiness, sadness, anger, surprise, disgust, fear)	7 (happiness, sadness, anger, surprise, fear, disgust and neutral)	7 (happiness, sadness, anger, surprise, fear and disgust and neutral)
Age	Good representation	Limited	Good representation
Ethnicities	Good representation	Limited	Good representation
Gender	Includes Male and Female	Includes Male and Female	Includes Male and Female
Lighting Condition	Varied throughout	Varied throughout	Varied throughout
Poses	Multiple	Multiple	Multiple

4.0 Methodology

For this project, the methodology I have chosen to use is the Cross-Industry Standard Process for Data Minin or also called CRISP-DM. CRISP-DM is a widely used methodology for data mining and analytics projects, providing a structured and iterative approach to address business objectives and analyse data. The life cycle of a data mining project is outlined in the CRISP-DM reference model for data mining. It includes a project's phases, as well as the tasks and outcomes associated with each phase.



(Hotz, 2023)

The life cycle of a data mining project is broken down in six phases which are shown in the above diagram.

The sequence of the phases is not strict. The arrows indicate only the most important and frequent dependencies between phases, but in a particular project, it depends on the outcome of each phase which phase, or which task of a phase, has to be performed next. (Wirth, n.d.). In the diagram, the outer circle represents the circular nature of data mining itself. Once a solution is implemented, data mining is not over. Lessons learned from the process and the implemented solution can lead to new, frequently more precise business queries. The lessons learned from earlier data mining processes will be applied to later iterations.

Other methodologies were potentially applicable to this project, and it’s needs. One such methodology is the Knowledge Discovery in Databases or KDD methodology. I decided against this methodology as it is often used in the context of large-scale data mining projects, which doesn’t apply to this project.

Here are the CRISP-DM Phases broken down, with relevance to this project:

1. **Business Understanding**
    - Understanding the project's goals, objectives, business context, and requirements for facial expression recognition and API comparison are the main points of this heading.
    - It entails determining the project's scope, identifying the key questions to be answered, and identifying the stakeholders.
    - The findings found in this stage are outlined in section (1) Introduction.

2. **Data Understanding**
    - This step involves exploring and understanding the data used in the project, which includes the labelled image datasets obtained from publicly available sources.
    - It includes data collection, data quality assessment, and initial data exploration to gain insights into the dataset's characteristics and structure.
    - It was fundamental that all image datasets contained images that were compatible with both API’s, contained associated metadata with labels associating emotions with the images, labels obtained and approved by a qualified body.
    - There were many other datasets that were explored to potentially be used in this project such as Affectnet (Mahoor, n.d.) and Japanese Female Facial Expression or JAFFE (Lyons, 2019), to name two. Some images were sampled from these datasets, but it was decided that they did not meet the criteria.
    - Therefore the three datasets: Cohen-Kanade or CK (Lucey, n.d.), Facial Expression Recognition or FER (al., 2019) and Expressions in the Wild or ExpW (al., 2015) were chosen.

3. **Data Preparation**
    - In this phase, the data is prepared for analysis by transforming, cleaning, and integrating the datasets to ensure they are suitable for comparison.
    - This step includes removing duplicates, handling missing data, formatting data, and preparing the data for further analysis.
    - There were several pre-processing steps taken before conducting an analysis with this data:

- The format of the images was checked, to be compatible with both API's the images had to be in JPEG or PNG format.
- Corrupt or incomplete image files were sought out and removed.
- Python code was written which opened the dataset and grouped the images into sub-folders based on their dominant emotion as outlined in the metadata.
- Python code was also written to create subsets of the image datasets to meet Luxand limitations. The datasets were divided as follows.
- CK contains 981 images, this whole dataset was used.
- ExpW contains 84.8k images. All images categorised as 'Neutral' for their dominant emotion were used (1048 images). The remaining images were split into random samples of 3500 images for each emotion (6 x 3500).
- FER was split into test and training data which contained 28.7k and 7.1k test images respectively. All images categorised as 'Disgust' were included as neither the test nor training set met the subset number. The remaining emotions were split into subsets of 3500 random samples. 2800 train images and 800 test to reflect the original 80:20 split.

4. Modelling

- This stage focuses on applying the chosen methodologies (in your case, comparing AWS Rekognition and Luxand FaceSDK) to the prepared data.
- It involves running the algorithms or APIs, evaluating their performance, and comparing the accuracy of the emotion detection capabilities.
- This is the stage where the APIs were called on the images and the emotion recognition took place. There were different methods for each API.

4.1 AWS Rekognition Modelling

- AWS region and profile name were set for authentication.
- AWS Rekognition and S3 client were initialised. AWS Rekognition for emotion detection and S3 for image storage.
- An S3 bucket was created and specified within the code to dedicate to this analysis to store the images.
- Output paths were allocated for the CSV and JSON files.
- An empty dictionary was created titled 'emotions' to store he detected emotions.
- A loop is set up to go through each image in the specified folder, upload it to the S3 bucket, use AWS Rekognition to detect emotions.
- The full AWS Rekognition response is written to a JSON file, and the emotion confidence levels is extracted and written to a CSV.

4.2 Luxand FaceSDK Modelling

- The Luxand API token is set.

- The image directory input is specified, along with the CSV and JSON output locations.
- A function is defined to use the Luxand API to detect emotions in an image, this blueprint is provided with the API token. (Luxand, 2023)
- It loops through the image directory file, for each image the function is called, and the emotion is detected.
- The full output is written back into a JSON file and the confidence levels associated with the emotions is written to a CSV.

5. Evaluation

- This step involves assessing and validating the results obtained from the modelling phase.
- It includes evaluating the accuracy of the emotion detection results against the original labelled dataset and comparing the performance of the two platforms.
- The CSV files were used primarily for the evaluation of performance.
- There were two statistics that were the main focus, the 'Mean Confidence' and the 'Binary Accuracy'.
- The 'Mean Confidence' was taken as an average of each emotion in the CSV file and is displayed in the 'Analysis' section of this report. This is a key indicator of the API's overall performance and also shows and interesting contrast when comparing the AWS Rekognition results and Luxand FaceSDK results. It also shows which performance are recognised with greater confidence and how this varies throughout the different datasets.
- The 'Binary Accuracy' was a measure followed issues surrounding the detection of a face in the Luxand API detection. The Binary Accuracy focuses on the images that were classified, and regardless of the confidence levels, how many of them had the highest confidence for the correct dominant emotion. These results can be found in the 'Results' Section of the project. It is also interesting to see that there is not much variance when comparing both metrics in the context of AWS Rekognition, but a variance is present when comparing it from Luxand.

6. Deployment

- This heading relates to the practical deployment of the findings and recommendations from the evaluation phase.
- It involves discussing the implications of the comparison results and providing recommendations for stakeholders when choosing a facial recognition API for their applications.
- There was no practical deployment in this project, but it has done a lot of background work to pose future research questions or to provoke further analysis into the topic.

5.0 Analysis

Cohn-Kanade Dataset

Average Confidence Levels by Emotion in Cohn-Kanade (CK) Dataset using AWS Rekognition								
	AWS RESULT							
CK LABEL	**HAPPY**	**SURPRISED**	**FEAR**	**SAD**	**ANGRY**	**DISGUSTED**	**CONFUSED**	**CALM**
HAPPY	94.37	8.27	6.07	2.23	0.44	1.37	0.49	0.70
SURPRISE	0.23	94.20	11.25	2.79	0.38	0.45	0.40	2.39
FEAR	0.81	14.10	24.54	44.72	5.03	24.89	1.05	6.26
SADNESS	0.59	7.11	6.58	77.01	9.95	14.19	1.88	7.11
ANGER	0.43	7.92	6.12	7.36	62.85	4.86	8.95	15.55
DISGUST	0.36	6.49	6.04	3.95	53.00	41.69	1.86	0.64
CONTEMPT	12.58	10.06	6.07	9.17	3.09	9.54	1.77	62.53

Figure 5.1 Average Confidence Levels by Emotion in Cohn-Kanade (CK) Dataset using AWS Rekognition

When searching for the emotion ‘Happy’ within the Cohn-Kanade dataset (SHAWON, 2018) , The highest percentage is for images pre-labelled as ‘Happy’ at a detection average of 94.37%. The other emotions have much lower percentages and Figure 5.1, showed its highest confidence levels when labelling images with this emotion.

For images in this dataset labelled as showing ‘Surprise’, AWS had a 94.2% detection rate on average. The dominant emotion in all these images was clearly labelled as ‘Surprise’. Despite listing ‘Fear’ as one of the emotions this API can detect, it had a higher average as labelling these image as ‘Sad’. The API showed high average confidence levels labelling images showing ‘Fear’, confusing it with 2 other emotions. These images were classified as having a dominant emotion of ‘Sad’ (44.7%), ‘Disgusted’ (24.9%) or ‘Fear’ (24.5%). It is also worth noting that 14% of images in this emotion subset were also labelled as having ‘Surprise’ as their dominant emotion.

The API showed a high average confidence when labelling pre-defined ‘Sad’ images as portraying ‘Sad’ as their dominant emotion at 77%. Similarly, for ‘Angry’ images, the API showed a high confidence level at 62.8%. An interesting take here is that the API labelled 15% of these images as exhibiting the emotion ‘Calm’. This is an interesting observation as some may argue calm and anger are opposite emotions.

There was a confidence level of 41.7% when labelling ‘Disgust’ images as having ‘Disgust’ as their dominant emotion. However, the API had a higher level of classifying these images as ‘Angry’ (53%). The images that were pre-labelled as exhibiting ‘Contempt’ were classified with a 62.5% confidence rate as exhibiting ‘Calm’. A state of inner tranquillity and emotional stability is typically associated with calmness, whereas contempt is a negative emotion characterized by scorn and disrespect for others. Contrary to contempt, which entails feelings of superiority and contempt, calmness is associated with relaxation and tranquillity.

The AWS API (Services, 2023) also classifies the emotions ‘Confused’ and ‘Calm’ which we did not have pre-defined images showing. It showed a reasonably low confidence when classifying images as ‘Confused’, the highest being 8.9% for images exhibiting ‘Anger’.

In this analysis, the API demonstrated strengths and weaknesses in accurately detecting and labelling emotions within the Cohn-Kanade dataset. The API performed well in identifying images labelled as 'Happy' and 'Surprise', achieving high detection rates of 94.37% and 94.2% respectively. However, it faced challenges when classifying images labelled as 'Fear', often misclassifying them as 'Sad', 'Disgusted', or 'Fear' itself. The API showed high confidence in labelling 'Sad' and 'Angry' images, but it also misclassified a portion of 'Angry' images as 'Calm'. Additionally, the API had difficulty distinguishing between 'Disgust' and 'Angry' expressions, often mislabelling 'Disgust' images as 'Angry'.

Notably, the API misclassified images labelled as exhibiting 'Contempt', instead classifying them with high confidence as 'Calm'. These observations indicate the API's strengths in recognizing certain emotions, but also highlight its weaknesses in accurately differentiating between closely related emotions and potential misclassifications.

Average Confidence Levels by Emotion in Cohn-Kanade (CK) Dataset using Luxand FaceSDK								
	LUX RESULT							
CK LABEL	ANGER	DISGUST	FEAR	HAPPINESS	SADNESS	SURPRISE	CONTEMPT	NEUTRAL
ANGER	43.43%	0.31%	0.03%	0.02%	9.27%	0.09%	13.15%	33.72%
DISGUST	10.86%	0.07%	0.01%	0.00%	1.85%	0.02%	0.87%	1.97%
FEAR	0.78%	0.05%	0.00%	0.01%	0.26%	0.03%	2.67%	10.78%
HAPPY	0.00%	0.05%	0.00%	0.01%	0.21%	0.00%	2.51%	10.50%
SADNESS	0.25%	0.19%	0.03%	0.00%	4.95%	0.07%	1.27%	1.49%
SURPRISE	0.01%	0.04%	0.00%	0.01%	1.02%	0.00%	2.23%	7.41%

Figure 5.2 Average Confidence Levels by Emotion in Cohn-Kanade (CK) Dataset using Luxand Face SDK

- It is evident from just a glace that the confidence levels of emotion detection are significantly lower for Luxand FaceSDK than for AWS Rekognition. The highest confidence level shown in the analysis of the CK dataset was 43.43%, which when looking at Figure 5.2, this would be considered a low confidence detection. The next highest confidence level is 13.15%, and then 10.86%. The confidence levels overall are significantly lower than those given by AWS Rekognition.
- However, the highest confidence level is found when matching a dominant emotion to an image is correctly representative of the image, as seen in figure 5.2. This indicated the strongest confidence exhibited by Luxand was when assigning Anger as a dominant emotion in an image.
- Another difference between this analysis and the Figure 5.1, is that there are some values at 0%. This was initially noticed in the testing phase, that if emotions were detected in an image, they were listed. However, if one emotion was not detected in the image, it would

not be mentioned in the output. Assuming this value as 0 was a step taken when processing the results.

In conclusion, the analysis highlights that Luxand FaceSDK exhibits lower confidence levels in emotion detection compared to AWS Rekognition. While it may accurately identify the dominant emotion in an image, the overall detection confidence is relatively low. Additionally, Luxand FaceSDK tends to miss certain emotions entirely, as evidenced by the 0% values. These findings suggest that AWS Rekognition may offer more robust and comprehensive emotion detection capabilities.

	% Failure
ANGER	86.03%
DISGUST	72.47%
FEAR	68.42%
HAPPY	84.13%
SADNESS	81.18%
SURPRISE	72.91%

Figure 5.3 Percentage of failure when detecting a face in Cohn-Kanade (CK) Dataset using Luxand Face SDK

An investigation into the rates of failure when detecting a face in the images of the Cohn-Kanade Dataset (SHAWON, 2018) The results can be seen above in Figure 5.3.

Initially, it was contemplated that this error rate could be attributed to the size of the images. All images used for this analysis are 48x48 pixels. A test was performed to increase the size of the image to investigate if this would make a difference to the face detection, but this resolution was unsuccessful.

It was also investigated that this error may lie within the code, so a sample of random images from the dataset were manually uploaded to the Luxand FaceSDK API demo (Luxand, 2023), but this did not resolve the issue either.

EXPW DATASET

Average Confidence Levels by Emotion in Expressions in the Wild (ExpW) Dataset using AWS Rekognition								
	AWS RESULT							
EXPW LABEL	**HAPPY**	**SURPRISED**	**FEAR**	**SAD**	**ANGRY**	**DISGUSTED**	**CONFUSED**	**CALM**
HAPPY	66.81	11.05	7.20	8.33	4.49	3.38	1.65	11.96
SURPRISE	3.97	53.28	19.68	7.16	14.40	3.26	3.85	10.40
FEAR	3.02	27.33	38.27	11.67	21.91	5.96	2.24	5.42
SADNESS	3.50	10.78	11.70	51.14	11.22	6.11	4.51	21.41
ANGER	1.76	12.06	9.77	7.80	65.69	5.82	2.58	9.79
DISGUST	3.34	11.10	9.82	20.35	21.21	9.10	7.99	34.61
NEUTRAL	4.99	11.32	7.83	21.36	11.44	5.01	3.93	51.52

Figure 5.4 Average Confidence Levels by Emotion in Expressions in the Wild (ExpW) Dataset using AWS Rekognition

Shown above in Figure 5.4 is the average confidence levels for emotion detection in the ExpW dataset as determined by AWS Rekognition. In this table, there is much lower confidence levels for classifying the dominant emotion than seen in the CK dataset. This could be because this dataset is significantly larger than CK so there is a larger margin for error, or the API could simply require more training on these images.

As seen in the first 5 emotions, the dominant image is correctly assigned. ‘Happy’ with 66.81% confidence, ‘Surprise’ with 53.28% confidence, ‘Fear’ with 38.27% confidence, ‘Sad’ with 51.14% confidence and ‘Angry’ with 65.69% confidence.

In the dataset, images labelled as showing "Disgust" as the dominant emotion were instead classified by the API as having "Calm" as the dominant emotion. The average confidence for the "Calm" classification was 34.61%, followed by "Angry" with an average confidence of 21.21%, and "Sad" with an average confidence of 20.35%. This misclassification suggests that there may be challenges in accurately identifying and distinguishing between these specific emotions. Certain emotions, such as happiness and surprise, appear relatively frequently across multiple rows, while others, like confusion and disgust, have lower frequencies.

It also worth noting that the number of images associated with ‘Neutral’ as a dominant emotion were significantly lower than images associated with any other emotion. This could be an indicator that the dataset creators themselves had issues finding images portraying such an emotion.

An analysis was not conducted on the Expressions in the Wild Dataset (al., 2015) using Luxand FaceSDK due to API limitations.

FER DATASET

TEST

Average Confidence Levels by Emotion in Facial Expression Recognition (FER) Test Dataset using AWS Rekognition								
	AWS RESULT							
FER LABEL	**HAPPY**	**SURPRISED**	**FEAR**	**SAD**	**ANGRY**	**DISGUSTED**	**CONFUSED**	**CALM**
HAPPY	67.79	20.26	7.58	3.77	2.60	5.30	0.86	6.62
SURPRISE	2.87	82.19	18.48	3.53	1.89	1.48	0.66	2.50
FEAR	2.16	19.21	29.06	21.58	14.22	7.77	4.31	19.24
SADNESS	1.56	11.28	10.23	44.58	8.09	8.57	3.31	32.08
ANGER	1.39	13.88	10.81	10.98	51.54	8.05	2.57	16.33
DISGUST	0.46	9.14	8.30	12.26	22.46	57.49	0.82	4.55
NEUTRAL	5.36	12.61	7.79	17.59	5.79	6.50	2.81	58.32

Figure 5.5 Average Confidence Levels by Emotion in Facial Expression Recognition (FER) Test Dataset using AWS Rekognition

From the beginning, the information shown in Figure 5.5 is supportive of the statement that AWS Rekognition can detect the dominant emotion portrayed in an image correctly, with average confidence levels as high as 82.19%. The API correctly classified 6 of 6 emotions correctly. It also classified 'Neutral' images as having a dominant emotion of 'Calm' which is the closest match possible with these labels, some may even argue calm and neutral could mean the same thing.

There is no particularly strange or peculiar outliers shown within these averages and based on this table alone, it shows that AWS Rekognition is fully capable of detecting the correct dominant emotion.

Average Confidence Levels by Emotion in Facial Expression Recognition (FER) Test Dataset using Luxand FaceSDK								
	LUX RESULT							
FER LABEL	ANGER	DISGUST	FEAR	HAPPINESS	SADNESS	SURPRISE	CONTEMPT	NEUTRAL
ANGRY	19.23%	1.39%	0.61%	1.97%	2.26%	2.72%	1.06%	16.85%
DISGUST	15.28%	17.66%	0.27%	10.53%	1.83%	0.42%	13.69%	18.09%
FEAR	2.83%	0.20%	10.43%	2.28%	6.86%	12.66%	0.71%	24.34%
HAPPY	0.00%	0.00%	0.12%	28.88%	0.01%	0.95%	0.02%	1.92%
SAD	0.77%	0.03%	0.35%	0.72%	11.98%	0.44%	0.53%	28.42%
SURPRISE	1.54%	0.07%	2.53%	6.48%	0.84%	55.94%	0.33%	3.33%

Figure 5.6 Average Confidence Levels by Emotion in Facial Expression Recognition (FER) Test Dataset using Luxand FaceSDK

Aside from large margins of 'failure to detect a face in the image', Luxand Face SDK is a highly rated API for emotion detection.

To exclude the failed recognition tests (outlined in pre-processing GIVE SECTION). This matrix focuses on the average confidence levels when a face is detected, and emotions are also detected. Again, the data in Figure 5.6 is much lower than that in Figure 5.5, proving that AWS has higher confidence levels for most emotions so far in this analysis.

An interesting to note is of the two emotion classifications so far in figure 5.2 and 6.6 , 'Angry' is the only emotion that has been correctly classified in both instances. 'Surprise' was one of three emotions correctly classified by Luxand in the FER Test dataset. This emotion had an average confidence of 55.94%, significantly higher than the 0% confidence assigning 'Surprised' as the dominant emotion exhibited with the CK dataset.

It is also worth noting for analysis, that the number of images that were classified as 'Disgust' being the dominant emotion was significantly lower than the other emotions.

Three image subsets 'Disgust', 'Fear' and 'Sad' have highest average confidence classifying them as portraying 'Neutral' as a dominant emotion. This is like another three emotions mis-attributed as 'Neutral' in the CK dataset. This may indicate that Luxand may have a habit to classify images as 'Neutral' due to a lack of training on other emotions.

	% Failure
ANGER	77.89%
HAPPY	76.60%
DISGUST	91.07%
FEAR	82.60%
NEUTRAL	75.18%
SAD	79.74%
SURPRISE	77.18%

Figure 5.7 Percentage of failure when detecting a face in Facial Expression Recognition (FER) Test Dataset using Luxand Face SDK

TRAIN

Average Confidence Levels by Emotion in Facial Expression Recognition (FER) Train Dataset using AWS Rekognition								
	AWS RESULT							
FER LABEL	HAPPY	SURPRISED	FEAR	SAD	ANGRY	DISGUSTED	CONFUSED	CALM
HAPPY	66.63	21.45	7.65	3.91	2.51	5.16	0.89	6.65
SURPRISE	2.53	80.16	19.97	3.24	2.06	1.31	0.61	3.26
FEAR	1.79	20.70	30.72	20.98	12.82	7.39	3.63	19.07
SAD	1.44	10.53	10.37	47.50	7.69	8.86	2.91	30.98
ANGRY	1.85	13.64	9.98	10.37	51.79	8.88	2.09	17.07
DISGUST	0.65	10.47	7.52	9.80	18.45	62.82	0.95	4.72
NEUTRAL	6.09	12.66	7.52	16.59	5.81	6.50	2.45	59.17

Figure 5.8 Average Confidence Levels by Emotion in Facial Expression Recognition (FER) Train Dataset using AWS Rekognition

Similar to Figure 5.5, AWS Rekognition correctly classifies all of the dominant emotions in the FER Train dataset. It also labels most 'Neutral' images as portraying 'Calm' as the dominant emotion, which as said above is a practical assumption. There are no major outliers or anything out of the ordinary in this analysis.

It is also worth noting for analysis, that the number of images that were classified as 'Disgust' being the dominant emotion was significantly lower than the other emotions. However, this does not appear to have had an effect on the result.

Average Confidence Levels by Emotion in Facial Expression Recognition (FER) Train Dataset using Luxand FaceSDK								
	LUX RESULT							
FER LABEL	ANGER	DISGUST	FEAR	HAPPINESS	SADNESS	SURPRISE	CONTEMPT	NEUTRAL
ANGER	21.13%	0.95%	0.73%	3.76%	0.29%	1.37%	0.01%	0.32%
DISGUST	15.85%	16.15%	4.73%	0.76%	9.51%	6.47%	14.93%	11.57%
FEAR	3.99%	0.44%	13.24%	2.04%	6.92%	13.21%	0.64%	18.58%
HAPPY	0.12%	0.00%	0.11%	31.64%	0.10%	1.83%	0.04%	2.04%
SADNESS	0.62%	0.34%	0.23%	0.87%	12.12%	0.74%	0.39%	27.29%
SURPRISE	0.70%	0.16%	1.73%	7.53%	0.42%	54.76%	0.08%	5.03%

Figure 5.9 Average Confidence Levels by Emotion in Facial Expression Recognition (FER) Train Dataset using Luxand FaceSDK

For the FER Train dataset, 4 of 6 emotions were classified correctly and there is no major outliers or peculiar metrics in this chart. The success shown with both AWS Rekognition and Luxand when dealing with this dataset could be indicative that the datasets may have been trained on such images or variations.

The metadata used for the FER dataset is FER-PLUS, which is an updated of the original metadata FER-2013 associated with the dataset. This would have been interesting to incorporate into the analysis but unfortunately this data could not be located.

Below in figure 5.10, the percentage failure in Luxand is shown, as shown all these failures are above 70%, which is not reliable for a resource. Although the predictions are correct, this margin of failure is something to consider that is associated with this API.

	% Failure
ANGER	77.94%
HAPPY	78.47%
DISGUST	91.76%
FEAR	83.22%
NEUTRAL	74.62%
SAD	82.11%
SURPRISE	75.58%

Figure 5.10 Percentage of failure when detecting a face in Facial Expression Recognition (FER) Train Dataset using Luxand Face SDK

6.0 Results

In the previous section, an analysis of emotion detection using two different APIs: AWS Rekognition and Luxand FaceSDK took place. The datasets used for evaluation included the Cohn-Kanade (CK) dataset, Expressions in the Wild (ExpW) dataset, and Facial Expression Recognition (FER) dataset. This analysis primarily focused on the average confidence levels for each emotion, dataset, and API.

In the analysis of the CK dataset, AWS Rekognition demonstrated high accuracy in detecting emotions such as 'Happy' and 'Surprise', with detection rates of 94.37% and 94.2% respectively. However, it struggled with correctly classifying 'Fear' images, often misclassifying them as 'Sad', 'Disgusted', or 'Fear' itself. The API also had difficulties distinguishing between 'Disgust' and 'Angry' expressions. Luxand FaceSDK, on the other hand, had lower confidence levels compared to AWS Rekognition for classifying dominant emotions. Luxand FaceSDK frequently misclassified images labelled as 'Disgust' as 'Calm'.

The ExpW dataset analysis showed that AWS Rekognition, achieved relatively higher accuracy in detecting dominant emotions across various subsets of the ExpW dataset.

The FER dataset analysis showed that both AWS Rekognition and Luxand FaceSDK achieved high accuracy in detecting the dominant emotions in the test and train subsets. AWS Rekognition consistently classified all emotions correctly, while Luxand FaceSDK showed some misclassifications and lower confidence levels.

Overall, AWS Rekognition demonstrated stronger performance and higher confidence levels in emotion detection across the analysed datasets. Luxand FaceSDK showed limitations in accurately detecting and classifying emotions, as well as higher failure rates in face detection. These findings suggest that AWS Rekognition may offer more robust and comprehensive emotion detection capabilities compared to Luxand FaceSDK.

MEAN CONFIDENCE VS. BINARY ACCURACY

MEAN CONFIDENCE vs. BINARY ACCURACY Cohn-Kanade (CK) Dataset using AWS Rekognition		
	MEAN CONFIDENCE	**BINARY ACCURACY**
HAPPY	94.37	97.58
SURPRISE	94.20	93.57
FEAR	24.54	20.00
SADNESS	77.01	77.38
ANGER	62.85	69.63
DISGUST	41.69	41.81
MEAN	65.78	66.66

Figure 6.1 Mean Confidence vs. Binary Accuracy CK Datasets using AWS Rekognition

An alternative means was also computed in the analysis of the three datasets, the first example of which is evident in Figure 6.1. The 'Mean Confidence' values refers to the average confidence value given by the API when correctly predicting the dominant emotion in an image. The Binary Accuracy was computed by investigating the frequency in which the average highest confidence level corresponded to the dominant emotion as per the labels provided in the metadata.

As evident in figure 6.1, Binary Accuracy had a higher score for 4 of 6 emotions and has a higher overall mean. Although it is worth noting that the difference between 'Mean Confidence' and 'Binary Accuracy' is not significant for any emotion, and they both even out, resulting in a mean difference of just 0.88.

MEAN CONFIDENCE vs. BINARY ACCURACY Cohn-Kanade (CK) Dataset using Luxand FaceSDK		
	MEAN CONFIDENCE	**BINARY ACCURACY**
ANGER	43.43%	22.22%
DISGUST	0.07%	39.58%
FEAR	0.00%	26.09%
HAPPY	0.01%	53.13%
SADNESS	4.95%	40.00%
SURPRISE	0.00%	58.21%
MEAN	8.08%	39.87%

Figure 6.2 Mean Confidence vs. Binary Accuracy CK Datasets using Luxand Face SDK

This table directly reflects the extremely low confidence levels shown in Figure 5.2 for Luxand's interpretation of this dataset. Anger had an extremely high confidence at 43.43% and is also the only emotion consistently correctly recognised as correct through all three Luxand analysis. This chart gives a good indication at the capabilities of Luxand, despite having low confidence levels, for 5 of 6 emotions Luxand had highest confidence levels for recognising the correct emotion as dominant.

MEAN CONFIDENCE vs. BINARY ACCURACY Expressions in the Wild (ExpW) Dataset using AWS Rekognition		
	MEAN CONFIDENCE	**BINARY ACCURACY**
HAPPY	66.81	67.22
SURPRISE	53.28	50.46
FEAR	38.27	36.48
SADNESS	51.14	50.36
ANGER	65.69	67.22
DISGUST	9.10	8.44
MEAN	47.38	46.70

Figure 6.3 Mean Confidence vs. Binary Accuracy ExpW Dataset using AWS Rekognition

Similar to AWS Rekognition's interpretation of the CK dataset (Figure 6.10), similar values are seen in both the 'Mean Confidence' levels and the 'Binary Accuracy' do not vary that much, the variation between the two means is less than 1.

Something that stood out from this table was the emotion 'Disgust'. Many AWS confidence levels exhibited throughout this project are relatively high for the corresponding emotion, but here the binary accuracy for predicting disgust is only 8.44. This means that 91.56% on emotions in this dataset were mis-classified as having another dominant emotion by AWS. This further solidifies the statement that AWS Rekognition could be facing difficulty classifying images with the dominant emotion 'Disgust'.

MEAN CONFIDENCE vs. BINARY ACCURACY Facial Expression Recognition (FER) Test Dataset using AWS Rekognition		
	MEAN CONFIDENCE	BINARY ACCURACY
HAPPY	67.79	53.44
SURPRISE	82.19	80.14
FEAR	29.06	25.29
SADNESS	44.58	43.06
ANGER	51.54	53.44
DISGUST	57.49	59.46
MEAN	55.44	52.47

Figure 6.4 Mean Confidence vs. Binary Accuracy FER Test Dataset using AWS Rekognition

Like previous comparisons, there is little variance between 'Mean Confidence' and 'Binary Accuracy' for AWS Rekognition's performance on this dataset. This is a good indicator that AWS Rekognition can provide useful insights and tagging the correct emotion in most contexts.

MEAN CONFIDENCE vs. BINARY ACCURACY Facial Expression Recognition (FER) Test Dataset using Luxand FaceSDK		
	MEAN CONFIDENCE	BINARY ACCURACY
ANGER	19.23%	73.38%
DISGUST	17.66%	44.44%
FEAR	10.43%	49.59%
HAPPY	28.88%	98.16%
SADNESS	11.98%	68.09%
SURPRISE	55.94%	88.05%
MEAN	24.02%	70.28%

Figure 6.5 Mean Confidence vs. Binary Accuracy FER Test Dataset using Luxand FaceSDK

Contrary to figure 6.4, figure 6.5 exhibits a much higher 'Binary Accuracy' than 'Mean Confidence'. This further demonstrates Luxands abilities that are not apparent through the Mean analysis. It can recognise the dominant emotion in most contexts for this particular dataset and there is over 45% in the difference between the confidence and accuracy levels which is extremely significant.

MEAN CONFIDENCE vs. BINARY ACCURACY Facial Expression Recognition (FER) Train Dataset using AWS Rekognition		
	MEAN CONFIDENCE	BINARY ACCURACY
HAPPY	66.63	69.87
SURPRISE	80.16	78.33
FEAR	30.72	26.80
SADNESS	47.50	45.90
ANGER	51.79	53.42
DISGUST	62.82	64.45
MEAN	56.60	56.46

Figure 6.6 Mean Confidence vs. Binary Accuracy FER Train Dataset using AWS Rekognition

Like the FER Test comparison in figure 6.4, there is little variation between accuracy and confidence for the FER Train dataset. Again, this is indicative of the reliability with AWS Rekognition,

MEAN CONFIDENCE vs. BINARY ACCURACY Facial Expression Recognition (FER) Train Dataset using Luxand FaceSDK		
	MEAN CONFIDENCE	BINARY ACCURACY
ANGER	21.13%	14.29%
DISGUST	16.15%	34.29%
FEAR	13.24%	55.22%
HAPPY	31.64%	96.84%
SADNESS	12.12%	68.80%
SURPRISE	54.76%	85.94%
MEAN	24.84%	59.23%

Figure 6.7 Mean Confidence vs. Binary Accuracy FER Train Dataset using Luxand FaceSDK

As shown in Figure 6.7, most emotions have a higher ‘Binary Accuracy’ level than ‘Mean Confidence’. Like the analysis of the FER Test dataset in figure 6.5, the mean of ‘Binary Accuracy’ is significantly higher at 35%.

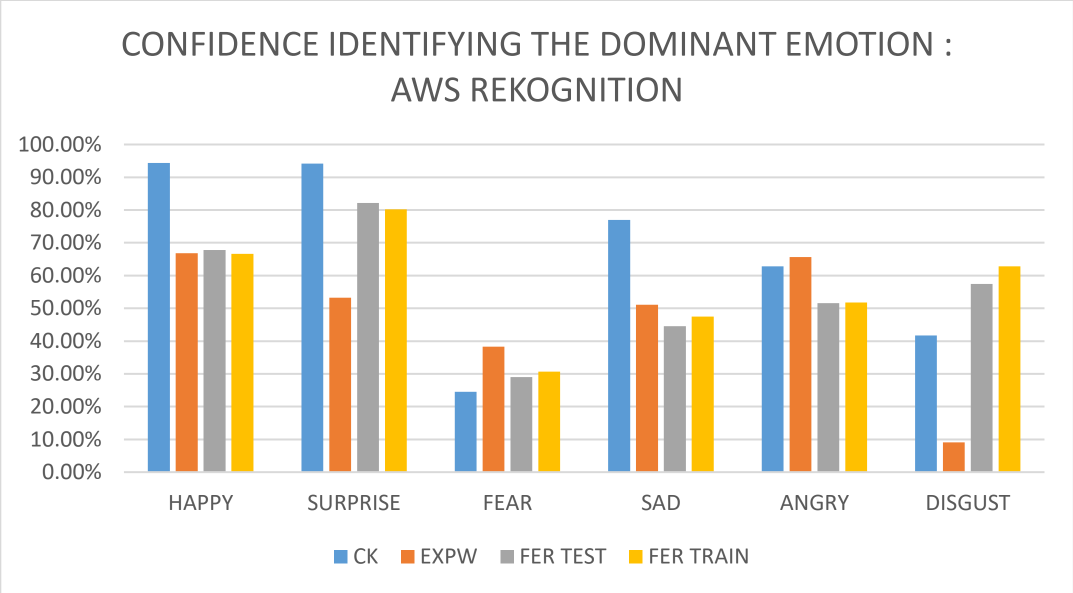


Figure 6.8 Confidence identifying the Dominant Emotion: AWS Rekognition

The highest levels of confidence with emotion belonged to the confidence levels associated with the CK dataset and the emotions ‘Happy’, ‘Surprise’ and ‘Sad’. There is no dataset that consistently has the highest identification, but also there is no dataset that consistently has the lowest.

The lowest confidence when identifying an emotion belongs to identifying ‘Disgust’ in the ExpW dataset. This was also pointed out in the analysis section of the paper.

‘Surprise’ was the emotion which received the highest confidence when identifying it correctly as the dominant emotion.

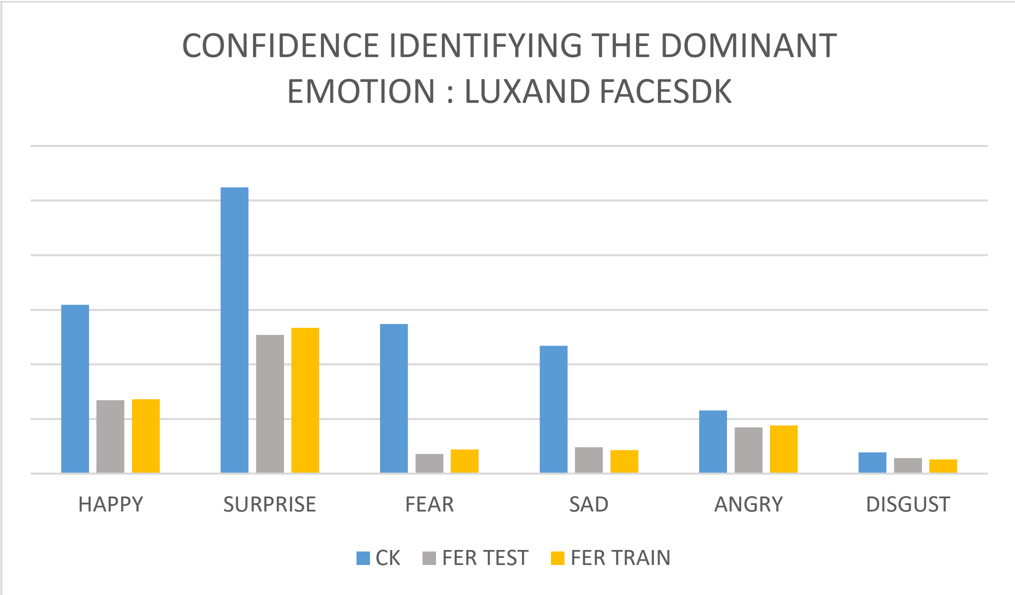


Figure 6.9 Confidence identifying the Dominant Emotion: Luxand FaceSDK

The highest levels of confidence with emotion belonged to the confidence levels associated with the CK dataset, just like in the AWS Rekognition stats and the emotions ‘Happy’ and ‘Surprise’. The Cohn-Kanade dataset consistently had the highest identification of the dominant emotion, while the Facial Expression Recognition dataset had lower levels in both the test and train set

The lowest confidence when identifying an emotion belongs to identifying ‘Disgust’ in the ExpW dataset. This was also pointed out in the analysis section of the paper.

‘Surprise’ was the emotion which received the highest confidence when identifying it correctly as the dominant emotion.

	CK		EXPW	FER TEST		FER TRAIN	
	AWS	**LUXAND**	**AWS**	**AWS**	**LUXAND**	**AWS**	**LUXAND**
HAPPY	Y	Y	Y	Y	Y	Y	Y
SURPRISE	Y	Y	Y	Y	Y	Y	Y
FEAR	N	Y	Y	Y	N	Y	N
SAD	Y	Y	Y	Y	N	Y	N
ANGRY	Y	Y	Y	Y	Y	Y	Y
DISGUST	N	Y	N	Y	N	Y	Y

Figure 6.10 Success Matrix to Determine Correct Emotion with the Highest Confidence

AWS Rekognition has a 75% success rate.

Luxand FaceSDK has a 62% success rate.

Luxand also has a average failure rate of 95% when detecting a face in an image.

COMPARE LUXAND SUCCESS RATE WITH % FAILURE TO GET STAT

7.0 Conclusions

Based on the comprehensive analysis conducted throughout this project, we have gained valuable insights into the performance and capabilities of two popular emotion detection APIs: AWS Rekognition and Luxand FaceSDK. The aim of this project was to evaluate and compare the emotion detection accuracy and confidence levels of these APIs using three different datasets: Cohn-Kanade (CK), Expressions in the Wild (ExpW), and Facial Expression Recognition (FER).

In the analysis of the CK dataset, AWS Rekognition demonstrated higher confidence levels and accuracy in correctly identifying dominant emotions, particularly for emotions such as 'Happy' and 'Surprise' with average confidence levels exceeding 90%. However, it struggled with accurately classifying 'Fear' images and distinguishing between 'Disgust' and 'Angry' expressions. On the other hand, Luxand FaceSDK exhibited lower confidence levels overall, with notable misclassifications of images labelled as 'Disgust' as 'Calm'.

The ExpW dataset analysis revealed that AWS Rekognition had a relatively higher accuracy in identifying the dominant emotions across different subsets. Though it had a lower binary accuracy rate for disgust than for other emotions, it had trouble correctly classifying images with that emotion as the dominant one.

In the evaluation of the FER dataset, both AWS Rekognition and Luxand FaceSDK demonstrated the ability to detect dominant emotions with high accuracy. AWS Rekognition consistently classified all emotions correctly, while Luxand FaceSDK showed some misclassifications and lower confidence levels.

The two APIs' performance was further highlighted by a comparison of mean confidence and binary accuracy. The reliability of AWS Rekognition in detecting emotions was generally demonstrated by a closer alignment between mean confidence and binary accuracy. While some emotions had higher binary accuracy rates despite having lower mean confidence levels, Luxand FaceSDK demonstrated variations between mean confidence and binary accuracy.

Overall, AWS Rekognition proved to be a robust and comprehensive emotion detection API, consistently achieving high accuracy rates and confidence levels across the analysed datasets. Its performance was particularly notable in correctly identifying emotions such as 'Happy' and 'Surprise'. On the contrary, Luxand FaceSDK exhibited lower confidence levels, limitations in accurately detecting and classifying emotions, and higher failure rates in face detection.

In conclusion, the aims of this project were successfully achieved by evaluating and comparing the performance of AWS Rekognition and Luxand FaceSDK in emotion detection. AWS Rekognition emerged as the stronger performer, offering more accurate and reliable emotion detection capabilities across the analysed datasets. However, further research and evaluation are necessary to explore additional factors such as dataset biases, training methodologies, and potential improvements to both APIs for more robust and comprehensive emotion detection in various real-world applications.

8.0 Further Developments or Research

Upon the completion of this project, there are several potential avenues for further developments and research in the field of emotion detection. Firstly, exploring ensemble approaches could be valuable, combining the strengths of multiple APIs such as AWS Rekognition and Luxand FaceSDK to improve overall accuracy and confidence in emotion detection. This could involve creating an ensemble model that leverages the predictions from multiple APIs to make more reliable emotion classifications.

Further research into how pre-processing methods like image enhancement or facial landmark normalization affect emotion detection algorithms may also be beneficial. Additionally, performing a more thorough analysis of dataset biases and their impact on emotion detection accuracy may reveal any biases that may be present in the analyzed datasets and help to address any training model limitations. Finally, investigating transfer learning techniques, where models pre-trained on massive datasets are adjusted or modified to improve emotion detection on particular datasets or domains, could be useful. Better generalization and improved performance on various datasets outside of those specifically analysed in this project would be possible as a result.

Ensemble Approaches: Building ensemble models by combining the predictions from multiple emotion detection APIs, such as AWS Rekognition and Luxand FaceSDK, can potentially improve the overall accuracy and confidence in emotion classification. Techniques like majority voting or weighted averaging can be employed to integrate the predictions and leverage the strengths of each API, resulting in more reliable emotion detection.

Pre-processing Techniques: Investigating various pre-processing methods, such as facial landmark normalization or image enhancement, can help emotion detection algorithms perform better. The visual features important for emotion recognition can be improved using image enhancement techniques like contrast adjustment or noise reduction. Face orientations or sizes can be normalized using facial landmark normalization techniques, which can reduce variations and make the data more consistent for improved emotion recognition.

Bias Analysis: Conducting a comprehensive analysis of dataset biases and their impact on emotion detection accuracy is crucial. This research direction involves investigating potential biases in the analysed datasets, such as demographic or cultural biases, and understanding their influence on emotion detection algorithms. Addressing these biases can help improve the fairness and generalizability of emotion detection systems.

Transfer Learning: Utilizing transfer learning strategies can improve emotion recognition abilities. Better generalization and better performance may result from pre-training models on expansive datasets like ImageNet or AffectNet and finetuning them on particular emotion recognition datasets. Models can learn from a variety of data sources using transfer learning, which enables them to recognize more intricate patterns and increase the precision of their emotion classification.

These developments and research directions can contribute to advancing the field of emotion detection by improving accuracy, addressing biases, and enhancing the robustness of emotion recognition algorithms. By exploring these avenues, researchers can further refine and optimize emotion detection systems for real-world applications in various domains, including affective computing, human-computer interaction, and social robotics.

9.0 References

al., B. e., 2019. *FER+ (Face Expression Recognition Plus dataset).* [Online]
Available at: https://paperswithcode.com/dataset/fer

al., Z. e., 2015. *ExpW (Expression in-the-Wild).* [Online]
Available at:
https://paperswithcode.com/dataset/expw#:~:text=The%20Expression%20in%2Dthe%2DWild,%E2%80%9D%2C%20or%20%E2%80%9Cneutral%E2%80%9D.

Bijlstra, G., n.d. *Example expressions from the Radboud Faces Database.* [Online]
Available at: https://www.researchgate.net/figure/Example-expressions-from-the-Radboud-Faces-Database-from-top-left-neutral-angry_fig1_331041447
[Accessed December 2022].

Bryant, D. & Howard, A., 2019. *A Comparative Analysis of Emotion-Detecting AI Systems with,* Hololulu: s.n.

Dolz, J., n.d. *Some Samples of the EmotiC Dataset.* [Online]
Available at: https://www.researchgate.net/figure/Some-Samples-of-the-EmotiC-Dataset_fig3_326312793
[Accessed December 2022].

Ekman, P., 2014. *Darwin's Claim of Universals in Facial Expression Not Challenged,* Berkeley, California: University of California, Berkeley.

Goldman, A. I. & SekharSripada, C., 2005. *Simulationist models of face-based emotion recognition.* Volume 94 Issue 3 ed. s.l.:Cognition.

Hotz, N., 2023. *What is CRISP DM?.* [Online]
Available at: https://www.datascience-pm.com/crisp-dm-2/

Lucey, P., n.d. *CK+ (Extended Cohn-Kanade dataset).* [Online]
Available at:
https://paperswithcode.com/dataset/ck#:~:text=CK%2B%20(Extended%20Cohn%2DKanade%20dataset)&text=sadness%2C%20and%20surprise.-,The%20CK%2B%20database%20is%20widely%20regarded%20as%20the%20most%20extensively,of%20facial%20expression%20classification%20metho
[Accessed December 2022].

Luxand, 2023. *Emotion recognition API Demo.* [Online]
Available at: https://luxand.cloud/emotion-recognition-api

Lyons, M. J., 2019. *JAFFE (Japanese Female Facial Expression).* [Online]
Available at: https://paperswithcode.com/dataset/jaffe

Mahoor, M., n.d. *AffectNet.* [Online]
Available at: http://mohammadmahoor.com/affectnet/

Malygina, M., Artemyev, M., Belyaev, A. & Perepelkina, O., 2019. *OVERVIEW OF THE ADVANCEMENTS IN AUTOMATIC EMOTION,* Moscow: Neurodata Lab.

Perkowitz, S., 2021. *The Bias in the Machine:,* s.l.: Charles Howard Candler Emeritus Professor of Physics, Emory University.

Raji, I. . D. & Buolamwini, J., 2019. *Actionable Auditing : Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products,* s.l.: AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.

Services, A. W., 2023. *Getting started with Amazon Rekognition.* [Online]
Available at: https://docs.aws.amazon.com/rekognition/latest/dg/getting-started.html
[Accessed 2022].

SHAWON, A., 2018. *Cohn-Kanada Dataset.* [Online]
Available at: https://www.kaggle.com/datasets/shawon10/ckplus
[Accessed December 2022].

Singh, R., Majumdar, P., Mittal, S. & Vatsa, M., 2022. *Anatomizing Bias in Facial Analysis,* s.l.: Proceedings of the AAAI Conference on Artificial Intelligence.

Wirth, R., n.d. *CRISP-DM: Towards a Standard Process Model for Data,* Ulm, Germany: DaimlerChrysler Research & Technology.

10.1Project Proposal

National College of Ireland

Project Proposal

"An Investigation into the Existence of Algorithmic Bias of Public Cloud Service Facial Recognition Algorithms"

27/10/2022

BSc. (Hons) Degree in Data Science

2023

Lauren Kelly

x19362533

x19362533@student.ncirl.ie

Contents

1.0 Objectives

Machine learning models that were trained on bias data depict the desired use cases incorrectly. As a result, the analysis's quality, precision, and dependability are poor. In this project, I will analyse three public cloud service facial recognition algorithms and apply them to the same dataset to investigate differences between the structure and output. While conducting this analysis, I will also be investigating algorithmic bias that has previously existed in these algorithms and the steps that were taken to eliminate this bias.

The goal of the project is to showcase an understanding of the algorithmic bias that exists today and measures that have been taken to minimise the impact this has on current and future AI machines.

2.0 Background

I chose to undertake this project as I read a book titled 'Invisible Women, exposing data bias in a world designed for men' by Caroline Criado Perez. This book highlights the important role that bias plays in data and the effects it can have if it is not dealt with accordingly. The book explores data bias through a gender-oriented view and left me with questions on other concepts which may exhibit this same bias. One that was particularly interesting to me was within facial recognition algorithms. The software created and used by companies and governments has now become such a standard feature in the average persons life, from unlocking phones to passport scanners to clocking in and out of work. Such a potent instrument, which can achieve accuracy scores as high as 99.97%, was not created overnight and had to overcome several obstacles. Many of these obstacles were oriented around bias within the data used to train the algorithm, thus resulting in bias within the finished product.

3.0 State of the Art

My primary focus will be on Public Cloud Service Facial Recognition algorithms: Google Cloud Vision API, AWS Rekognition and Azure Face API.

There have been many analyses carried out on bias within facial recognition algorithms and individually of all three of my chosen algorithms. These various analyses focus on the performance, detection accuracy and bias within the algorithm. I will conduct a unique analysis using the findings of these existing analyses and compare it with the reformation and adaptation of the algorithms. I will compare the three most popular available systems, which will differ from similar analysis online.

4.0 Data

For this work, I will be using publicly available data about the algorithms and their adaptations. I will cite reports internally and externally from the companies in relation to any existence of bias within previous versions of the algorithms. I will gain an understanding of the similarities and differences of the three chosen algorithms and choose a public dataset on which I can test and evaluate the suitability of each of the algorithms.

I will be using data relating to the history of each of the algorithms and study how the code differs form one another. Following this study, I will apply the publicly available code of each of the algorithms to a chosen public dataset and investigate how the operations and results vary.

5.0 Methodology & Analysis

The methodology I have chosen to follow for my project is the KDD methodology. In an ongoing process called Knowledge Discovery in Databases (KDD), evaluation metrics may be improved, mining can be honed, new data can be added, and results can be changed to provide new and more relevant outcomes.

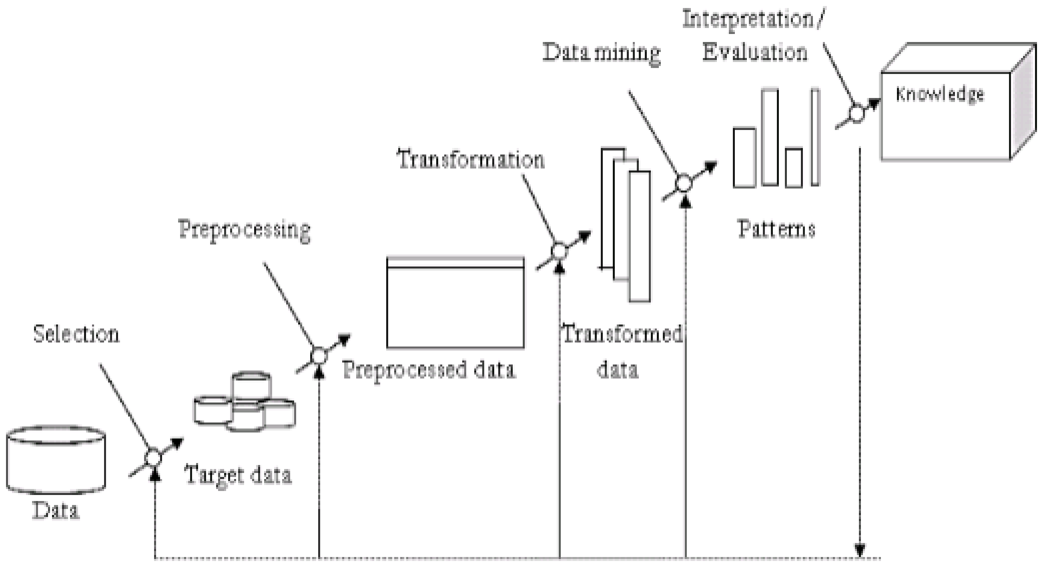


Fig KDD process

I will closely follow the project plan set out in part 7.0 of this document. Following this plan accompanied with frequent feedback from my supervisor will ensure that all requirements will be met. I will set out tasks, activities and milestones for each stage and then review them once a stage has been completed.

6.0 Technical Details

As part of this project, I must gain an understanding into the operation of each of the algorithms and also an understanding of the history and adaptations that took place with the aim of eliminating algorithmic bias.

I will also be incorporating any press / newspaper reports regarding any backlash received when using the various algorithms. It is important to ensure any of the articles I will include are from a trustworthy source and are relevant to the task of the project.

I will also ensure the chosen dataset on which I will preform the algorithms is suitable for each and not tailored to one specifically. This is crucial when making an analysis across all three algorithms as initially they could have been created for different purposes.

7. Project Plan

06/11/2022

Deep dive into Google Cloud Vision API.

- Find similar papers studying the facial recognition algorithms.
- Investigate any controversy around the algorithm regarding bias, investigate how the algorithm was adapted to overcome this problem.
- Create an approximate timeline for changes made and innovation added within the algorithm.
- Find publicly available code to test at a later stage.

13/11/2022

Deep dive into AWS Rekognition

- Find similar papers studying the facial recognition algorithms.
- Investigate any controversy around the algorithm regarding bias, investigate how the algorithm was adapted to overcome this problem.
- Create an approximate timeline for changes made and innovation added within the algorithm.
- Find publicly available code to test at a later stage.

20/11/2022

Deep dive into Azure Face API

- Find similar papers studying the facial recognition algorithms.
- Investigate any controversy around the algorithm regarding bias, investigate how the algorithm was adapted to overcome this problem.
- Create an approximate timeline for changes made and innovation added within the algorithm.
- Find publicly available code to test at a later stage.

27/11/2022

Compare deep dives similarities and differences. Investigate what led to different controversies raised with the three chosen hosts; AWS, Google and Azure.

04/12/2022

Compare elements of each algorithm: accuracy, underlying bias, performance and cost. Compare how each of the algorithm applies to the chosen dataset.

11/12/2022

Write up findings so far, re-evaluate to do list and asses any gaps in findings so far.

18/12/2022

Add in relevant code, graphs and cite relevant other work. Meet with project supervisor for any feedback to be incorporated before submission.

20/12/2022

Incorporate relevant feedback and finalise details for midpoint review submission.

10.1 Reflective Journals

JOURNAL 1

- Do not update main parts of the proposal, as you will detail them in the main part of the report --> Only add the GanttChart in section 7 if time allows.
- Comparison tables were not done or are incomplete (i.e., summary of research papers, datasets, services/APIs).
- Many important details are still not defined

- Report Feedback:

- Background: Add supporting references.

- Aims: expand the aims with more specific details (i.e., evaluate their performance and bias on age (children vs. adult), gender, race, basic emotions).

- Remove platforms as you will detail them later.

- Technologies: Mention that you will use Python, APIs, data extracted in JSON.

- SOTA: Remove paper titles, keep citation and add a brief summary paragraph on each academic paper. You need more academic papers (to be added later if not feasible by midpoint).Expand the services/APIs summary table with more columns (i.e., services) and rows (i.e., comparison criteria).

- Data: Add summary with comparison of datasets (e.g., size in MB, num images, file type, emotions, other labels like gender, age and race).

- Methodology: Add subsections for each step of KDD (or CRISP-DM if you plan to do deployment / dashboard). Add more details in each subsection.

- Analysis: Include more specific details about DS methods to be used.

- Results: You can break them in separate subsections for the different services / APIs. Include screenshots showing that you tested them (i.e., uploaded an image, and got the json response). You can also add summary table with the main information returned in the json file. Add text discussion and interpretation of the preliminary results/testing.

- Add meaningful captions for tables and figures.

- Include citation in the captions for figures you took from other sources (e.g.,Google pricing model)

- Check and follow the midpoint submission guidelines.

JOURNAL 2

- Started to implement Python code for AWS Rekognition --> Try to finalise it asap
- Contacted authors to get access to datasets (all 5 mentioned in the midpoint)
- Got access and downloaded AffectNet and EMOTIC as zip --> But cannot open them as they are too large
- Double check the links if they are to folders and if possible to download part of the datasets
- During next meeting show me the links and e-mails for all datasets
  You will have to add some novelty and contributions as compared to prior research (e.g., more/different APIs and datasets)

JOURNAL 3

- Downloaded a few datasets
- ExpW - Downloaded fine
- FERPlus - has original and new annotations, but pictures are too small 48x48 pixels, AWS Rekognition throws errors
- EMOTIC - has link but cannot open the zip as it is too big
- AffectNet – Unnavailable
- CK+ and JAFFE - Found link on GitHub
- Code: Finished test code for AWS with sample image.
- Update code to test image and save loccally on your PC the JSON response with the same file name as the image and platform (e.g., image_name_aws_rekognition_.json).
  JSON files need to be persisted to enable more in-depth analysis later on (will create separate code for that).

JOURNAL 4

- Focus on finalising the code for AWS Rekognition
- ExpW images are 48x48 pixels --> Gets labels through the API
- When trying to upload manually to the demo in the web console, it shows an error that pictures need to be larger than 80x80
- In NCI account gets AccessDeniedException --> Filled the IT request form for AWS resources
  Rescaled the image to 96x96 and uploaded to personal AWS account
- Did the Facial Analysis --> Saved the JSON manually
- Figure out how to so this and save the JSON automatically with Python
- Do not resize the images if they work OK through code
- Use your free tier account, if you face any limitation hopefully you get access through the NCI account soon

JOURNAL 4

- AWS: full access through NCI account, test all images of the datasets
- Luxand working fine, 50k monthly limit with basic trial plan
- Azure: could not connect properly, limitations too small
- Kairos: could not get access, NCI account not recognised as student e-mail, no reply from company
- Focus on AWS and Luxand and test maximum number of images possible
  Datasets:
- CK+ has 981 images --> test all images through both AWS and Luxand
- ExpW has 84.8k images --> AWS test all images; Luxand: all neutral (1048) + 6 * 3500 random sample for each other emotion (total 22k)
- FER+ has 28.7k train + 7.1k test images --> AWS test all images; Luxand all disgust + 7 x 3500 random sample for each emotion (2800 train + 700 test to reflect original 80:20 split).

- Organise all code and data files to be in 1 folder with subfolders
- Fix the part to save the json files (at the moment it is not working) --> Save json files in subfolders for different APIs and datasets
- Append AWS and luxand to file names to avoid risk of overwriting (e.g., <imagename>_aws.json)"
- Focus on tasks in the below order:
  1) AWS: test all images from the 3 datasets, save json files, extract emotions, compute accuracy values
  2) Luxand: test CK+, take random samples and test FER+, then ExpW

JOURNAL 5

- Computed confusion matrices based on average confidence
- There was an issue for Luxand --> Changed the formula to exclude 0% for failed images --> Mean confidence increased
- Added binary classification (i.e., assigned 1 if max confidence corresponded to the same emotion, 0 otherwise)
- Added new table comparing mean confidence vs. binary accuracy side by side
- For Luxand ran out of credits (did CK and FER, but not ExpW) --> The web dashboard still shows 50k balance, but API throws error.

Tasks:

1) Finalise analysis, improve and add all tables based on the feedback --> Copy paste them as proper tables not screenshots in the report.
2) Write a very good report to reflect the quality, novelty and complexity of the work (and maximise your grades).
3) Do the other deliverables and submit.

- Upload the CSV files generated by the Python code
- DO NOT upload the JSON files
- DO NOT upload the Excel spreadsheets with formulas

- I will follow up if you have the time and want to help writing a research paper based on the work.