POLITECNICO DI TORINO

Master's Degree in COMPUTER ENGINEERING

Master's Degree Thesis

Emotion Detection and Recognition

Supervisors

Prof. GABRIELLA OLMO

Prof. VITO DE FEO

Candidate

MOSTAFA DASHTI

JULY 2021

Summary

The aim of this graduation thesis, Emotion Detection and Recognition is to implement a cloud application that helps the patients of a rehabilitation clinic to do a particular test (GERT), through a web browser. Alexithymia is a subclinical disorder characterized by a decreased perception of the emotional states of a person that has a significant influence on mental health and social contact. Although this condition's clinical importance is high, the affected anterior insula (AI) contributing to alexithymia is uncertain. The word "Acquired Alexithymia" applies to alexithymia, which is demonstrated by a significant number of people with acquired brain injuries such as stroke or traumatic brain injury caused by an accident.
This test is typically part of the recovery process in a controlled environment set up by clinical experts. The structure and results of an application that gives patients the opportunity to do the test from home and get proper care are presented in this report. To make the application accessible in any place even at home, the author decided to use the cloud environment as infrastructure that allows developing a scalable application that needs a lot of computational resources to serve complex services to many patients.
In intelligent health contexts, an artificial intelligence-based application is designed as a solution for the detection and recognition of emotions. The aim of this application is to assess the emotional condition of the patients through the analysis of their facial expressions. Using labeled images from the ImageNet dataset, we initially trained a convolution neural network model to classify images recorded from patients into 7 different groups of emotions, including anger, happiness, sadness, disgust, fear, surprise, and natural, to achieve this goal. Then we improved our approach to estimate the emotion of the patient in continuous space in terms of two values. One for "valence" and one for "arousal" respectively. The new model used labeled images from the AffectNet dataset containing approximately 420K images that were labeled manually with two arousal and valence values. The valence shows the stimuli' pleasantness and is defined along the horizontal axis from the negative (unpleasant) pole to the positive pole (pleasant). On the other hand, Arousal refers to the intensity of emotion that is expressed by the same spectrum as valence along the vertical axis.

This graduation thesis is composed of five chapters. Each of them dealing with a different aspect of the implementation of the application. Chapter 1 is introductory and describes the problem and related solution based on facial emotion expression. Chapter 2 discusses the background information and theories behind cognitive science and facial emotion detection and recognition. Chapter 3 deals with cloud infrastructure and particularly the Google Cloud environment. Chapter 4 concentrates on problem resulting from pre-processing data that is obtained by a group of individuals and the classification model. The Conclusion is represented in Chapter 5. The main aim of the graduation thesis has been reached in this chapter.

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of support and assistance.
I would first like to thank my supervisors, Professor Gabriella Olmo and Professor Vito De Feo, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

"Mostafa Dashti
Turin 2021"

Table of Contents

List of Tables

List of Figures

Acronyms

AI
Artificial Intelligence

CNN
Convolutional Neural Network

CPM
Component Process Model

EEG
ElectroEncephaloGram

GP
Gaussian Process

GSR
Galvanic Skin Response

HMM
Hidden Markov Model

LSTM
Long Short-Term Memory

MSE
Mean-Square Error

PAD
Pleasure-Arousal-Dominance

PCA

Principal Component Analysis

RNN

Recurrent Neural Network

SD

Standard Deviation

SVM

Support Vector Machine

SVR

Support Vector Regression

VA

Valence-Arousal

Chapter 1

MACHINE LEARNING and HEALTH INDUSTRY

1.1 Mental Disorders

There are many various mental disorders, with several presentations. It is generally defined by a combination of abnormal thoughts, perceptions, emotions, reactions, and relations with others. Mental disorders include schizophrenia and other psychoses, depression, bipolar disorder, madness, developmental disorders including autism, and other disorders related to emotion recognition like alexithymia.

Health organizations have not yet enough reacted to the big duty of mental disorders. Therefore, the gap between the requirement for treatment and its supply is large globally. In many countries, between 76% and 85% of people with mental disorders do not get any remedy for their sickness.[1]
There are practical treatments for mental disorders and procedures to decrease their pain and distress. Access to health services that are prepared to provide therapy and social support is essential. There are also useful strategies for diagnosing mental disorders such as alexithymia and in this experiment, the author developed an AI-based application to help the patients to get checked from home.

1.1.1 Acquired Alexithymia

As Hogeveen et al., explain "Alexithymia is a subclinical condition characterized by impaired awareness of one's emotional states, which has profound effects on mental health and social interaction. Despite the clinical significance of this condition, the neurocognitive impairment(s) that lead to alexithymia remain unclear. Recent

theoretical models suggest that impaired anterior insula functioning might be involved in alexithymia, but conclusive evidence for this hypothesis is lacking."[2]
Recent theoretical activity recommends that alexithymia may be the result of a disrupted "interoception," which refers to a person's sensitivity to the internal state of their body [3]. According to these theories, interception involves the continuous production of models that predict the internal state of the body that are tested against input sensory evidence.
Differences between predictions and evidence are known as "prediction errors", which are hierarchically organized, and low-level errors represent basic somatosensory signals, that are used to update higher-level models.
At the highest level of the auditory prediction hierarchy, the anterior insula and the anterior cortex of the anterior contraction (ACC) are believed to represent our conscious emotional life. particularly, the anterior insula has a clear functional role in this context: by integrating prediction errors at the lower levels of the hierarchy, it creates awareness of one's feelings.[2]

1.1.2 GERT Test

GERT Definition

The ability to precisely explain the emotional expressions of others on the face, voice, and body is one of the most important components of successful social functioning and has been exposed to predict more reliable results in life.[4]

Existing tests to estimate the ability to identify the emotional states of others (emotions recognize ability [ERA]) are more focused on a single state (usually the face) and include only a limited number of emotions and restrict their ecological validity. In addition, their reliability is often inadequate.[5]

The Idea is to develop a new version of GERT test to improve results based on different sensors. In this research we tried to use Facial Emotion Expression, Speech Emotion Recognition, Galvanic Skin Response, EEG and ECG signals. In this thesis we focused on Facial Emotion Recognition, which is more reliable in compare to other methods.

GERT Function

The Geneva Emotion Detection Test (GERT; Schlegel, Grandjean, & Scherer, 2014) is a performance-based test to estimate individual variances in the ability of individuals to identify the emotions of others in the face, voice, and body. This

ability is one of the main elements of emotional competence or intelligence.

This includes short video clips in which ten actors (five men, five women) express 14 different emotions. These clips are taken from the Geneva Multimodal Emotion Portrayals Dataset (GEMEP, Bgernziger, et al., 2011). After each clip, participants are requested to decide which of the 14 emotions is exposed by the character.

Figure 1.1: After each video, 14 emotion words are presented, arranged in a circle that will help patient to rapidly select the appropriate emotion[6]

1.2 Machine Learning

"Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from data-mining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests."[7]

Machine learning is a subcategory of artificial intelligence (AI) that focuses on developing applications that learn from data and improve their accuracy over time without planning to do so. In data science, algorithms are sequences of statistical processing steps. In machine learning, algorithms are "trained" to find patterns

and features in large volumes of data for decision-making and prediction based on new data. The better the algorithm, the more accurate the data processing, decisions, and predictions become.

Nowadays, there are many examples of machine learning around us. In response to our voice commands, digital assistants control your smartphones, search the web and play music. based on what we have already bought, watched, or listened to, websites recommend products, movies, and songs. As we do our important works, robots do our regular works. Spam trackers stop spam from occupying our inboxes. Medical image analysis systems help doctors diagnose tumors that may have been missed during examinations. And the earliest self-driving cars move passengers.

More can be expected. As big data grows and computers become more powerful and cost-effective, and as scientists continue to develop more powerful algorithms, machine learning becomes more and more efficient in our life.

1.2.1 Artificial intelligence, machine learning, and deep learning

To understand the relationship between artificial intelligence (AI), machine learning, and deep learning these definitions can help:

Think of AI as the whole world of computing technology that displays anything remotely like human intelligence. Artificial intelligence systems can include anything from an expert system - a program that makes decisions based on complicated rules or logic - to a robot that cleverly develops the emotions of a human being, free will, and intelligence.

Machine learning is a subset of an AI program that teaches on its own. In fact, by digesting more data, it reprograms itself to perform the specific task it was designed to perform more accurately.

Deep learning is a subset of a machine learning program that teaches on its own, to do certain tasks with increasing accuracy without human intervention.

1.2.2 How machine learning works

There are four basic steps to building a model or machine learning program. This is usually done by data scientists who work closely with the business experts for

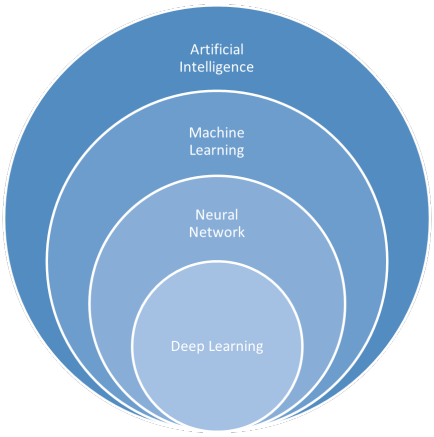

Figure 1.2: The relation between Artificial intelligence, machine learning, neural network, and deep learning

whom the model is being developed.

Training data set

Training data is a set of data that is used by the machine learning model to solve a problem designed to solve it. Sometimes, training data is annotated with data-"tagged" to indicate the features and classifications that the model needs to know. Other data is unlabeled, and the model must extract and classify those features alone.

In any case, training data must be properly prepared - be random and checked for biases or imbalances that may affect training. It should also be divided into two subsets: the training subset, which is used to teach the program, and the evaluation subset, which is used to test and modify it.

Learning Algorithm

The algorithm is a collection of statistical processing actions on data. The type of algorithm depends on the characteristic (labeled or unlabeled) and the quantity of data in the training datasets and the type of problem to be resolved.

Popular types of machine learning algorithms that are using labeled data include the following:

Sample-based algorithms: A good example of sample-based algorithms is K-Nearest Neighbor or k-nn. Uses classification to estimate the probability of membership of a data point in one group or another based on closeness to other data points.

Regression Algorithms: logistic regression and Linear regression are instances of regression algorithms used to learn relationships in data. Linear regression is utilized to predict the value of a dependent variable from the value of an independent variable. Logistic regression is getting used when the dependent variable is binary: A or B. For instance, a linear regression algorithm can be trained to guess a vendor's annual sales based on its relationship to seller's education or Years of experience. Support vector machine is another type of regression algorithm that is useful when it is more difficult to classify dependent variables.

Decision Trees: these algorithms use classified data to make suggestions according to a set of decision rules. For instance, a decision tree that advises betting on a particular team to win, place, or show can use team-related data (e.g., player's age, winning percentage, financial data) and enforces rules for these factors to suggest a decision.

There are also several algorithms working based on unlabeled data:

Neural Networks: A neural network is an algorithm that represents a network of computational layers with an input layer, where data is consumed; The network has one hidden layer, where the computations are performed, represents different results about the input. And an output layer where a probability is assigned to each conclusion. When the network has several hidden layers, each of which sequentially improves the results of the previous layer, we call it a deep neural network.

Clustering Algorithms: Consider clusters as a group. Clustering concentrates on distinguishing groups of similar entities and labelling them according to the group they belong to. It happens without any knowledge of the groups and their features. Types of clustering algorithms include Two-step, K-means, and Kohonen.

Association Algorithms: These types of algorithms find patterns and relationships in data and recognize association rules. These rules are comparable to the rules adopted in data mining.

Training and Model Creation

Algorithm training is a repeated process - it includes executing variables within the algorithm, comparing the output with the required results, modifying the weight and bias in the algorithm that may have a more reliable result, and running the variables repeatedly to the algorithm to get the best possible result. The resulting well-trained and accurate algorithm is the model - a critical contrast to remember because the "algorithm" and the "model" are mistakenly used correspondingly.

Model Improvement

The last step is to apply new data to the model and at best to improve its accuracy and performance over time. The source that the new data reaches from, depends on the solution and the problem that is being solved. For instance, a model built to detect spam checks email messages, while another model that drives a self-drive car transmits data from real-world interaction with the environment.

1.2.3 Machine learning Methods

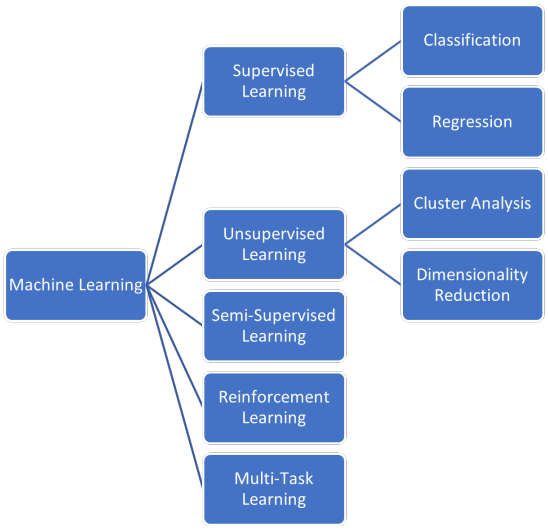

Figure 1.3: Machine learning methods hierarchy graph

The figure 1.3 shows the different categories of machine learning methods and their usage.

Supervised Learning

Supervised learning is a technique by which you can utilize labelled data for functional training, which you can then generalize to new instances. This training process includes a reviewer who can determine when the function is accurate or not, and then modify the function to get the correct result. Classic models like convolution neural networks usually trained by back-propagation algorithms, but there are many algorithms available.

Supervised learning process includes creating a model using labelled learning data that consists of input data and a desired output. The Supervising process is in the form of desired output, which in turn allows you to modify the model based on actual output. Once trained, you can use this model to new inputs to create a predictor or classifier that responds ideally.

Unsupervised Learning

Unsupervised learning uses unlabeled data and in real time, utilizes different algorithms to extract the important features needed to label, describe, and classify data, without human interference. Unsupervised learning is more about identifying patterns and connections in the data that people would miss. Consider spam detection, for instance, an unsupervised learning method can check enormous volumes of emails and discover the characteristics and patterns that show spam, and you will get better at labeling spam over time.

Semi-Supervised Learning

Semi-supervised learning provides a nice environment between supervised and unsupervised learning. Training process uses a smaller labeled datasets to supervise the classification and extraction of features from a more general, unlabeled datasets. Semi-supervised learning can solve the problem of not having sufficiently labeled data (or the ability to provide sufficiently labeled data) to train models based on supervised learning.

Reinforcement Learning

Reinforcement learning is not just a remarkable learning model, with the power to find out a way to map an input to an output, but also in some models like Markov decision processes, the model is able to map a set of inputs to dependent

outputs. Reinforcement learning exists in the context of situations (states) in an environment and possible moves (actions) in each situation. While the learning is in progress, the algorithm discovers the state-action pair in some conditions randomly to build a state-action pair table. Then within the practice of the learned information exploits the state–action pair rewards to settle on the easiest action for a given state that cause some goal state.

Multi-Task Learning

"Multi-Task Learning (MTL) is a sub-field of machine learning in which multiple tasks are simultaneously learned by a shared model. Such approaches offer advantages like improved data efficiency, reduced over-fitting through shared representations, and fast learning by leveraging auxiliary information."[8]

Machine learning, which uses useful knowledge in traditional data and uses information to support analyze future data, usually requires a large amount of labeled data to train a good model. A typical category of models in machine learning are deep learning models, that are many hidden layers as well as many parameters in the form of neural networks. These models typically require millions of data samples to learn the exact parameters.

However, some area, such as medical image analysis, may not be able to meet this need because they require more manual work to label the sample data. In these cases, multi task learning (MTL) [9], using useful information from other related learning, is a good model to help reduce the problem of data scatter.

Chapter 2

Literature Review

2.1 Cognitive Science

The interdisciplinary study of mind and intellect, which includes philosophy, psychology, artificial intelligence, neuroscience, linguistics and anthropology, is cognitive science. Its conceptual origins deep in the mid-1950s, when researchers started to establish mind theories based on complex representations and computational techniques in several areas.[10]

When the Cognitive Science Society was founded and the journal Cognitive Science started its work, its organizational origins were in the mid-1970s. Since then, over one hundred universities have developed cognitive science programs around the world, and many others have implemented cognitive science courses.

2.1.1 History

The history of study about the mind and its functions goes back to Ancient Greeks. When intellectuals such as Plato and Aristotle attempted to clarify the meaning of human intelligence. Until the nineteenth century, when experimental psychology grew, the study of mind stayed the province of philosophy. Behaviorists such as J. B. Watson believed psychology should limit itself to the study of the relationship between observable stimuli and observable behavioral responses. From respectable scientific discussion, talk of consciousness and mental models was banished.[12]

Behaviorism dominated the psychological scene through the 1950s, especially in North America. Around 1956, George Miller proposed that memory constraints could be overcome by re-encoding information piece-wise, mental representations involving mental methods for encrypting and decoding data. Early computers were

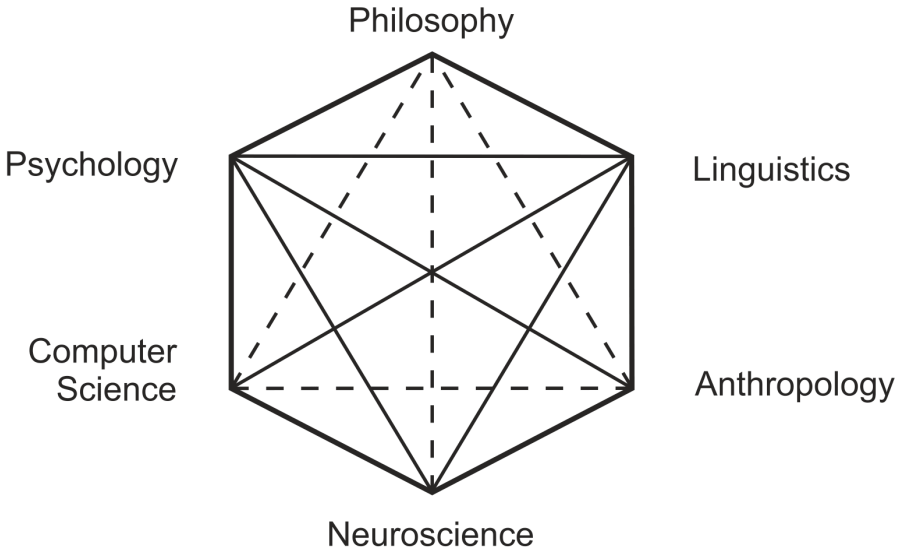

Figure 2.1: The fields that contributed to the birth of cognitive science [11]

only a few years old at the time, but engineers such as John McCarthy and Herbert Simon developed the field of artificial intelligence. Moreover, Noam Chomsky rejected behavioral assumptions about language as a learned habit and suggested explaining language understanding in terms of rules-based mental grammar instead.[13]

2.1.2 Methods

The cognitive sciences have integrated theoretical ideas, but we should appreciate the diversity of perspectives and methods that researchers in various fields contribute to study the mind and intelligence. Although cognitive psychologists today usually focus on computational theorizing and modeling, their main method is to research with human participants. People who frequently have undergraduate course needs are brought to the lab so that different types of thinking can be considered under controlled situations. For instance, psychologists have analytically examined the types of misunderstandings people make in deductive argumentation, the ways in which concepts are formed and applied, the speed with which people think with mental images, and the performance of people in solving problems using metaphors. Our summing-up about how the mind works should be more than "common sense" and introspection, as these can provide a misleading view of mental

procedures, many of them are not available consciously.[14]

Psychologists are increasingly attracting experimental participants from a variety of cultural sources. Psychological research that carefully approach mental procedures from different angles are crucial to the scientific nature of the cognitive sciences. Moreover, experimentation is a method used in experimental philosophy.

"In science, experiment without theory is blind, but theory without experiment is empty."[15] To discuss fundamental questions about the nature of the mind, psychological experiments must be interpreted in a theoretical framework that assumes representations and mental procedures. One of the most reliable ways to develop theoretical frameworks is to form and test computational models that are like mental operations. Researchers have developed computational models that mimic aspects of human performance to complete concept formation, mental imagery, psychological experiments on deductive reasoning, and similar problem-solving. Designing, building, and testing computational models are the primary method of artificial intelligence (AI), the category of computer science that deals with intelligent systems.

Similar to cognitive psychologists, neuroscientists usually perform controlled experiments, but their perceptions are very different because neuroscientists are directly related to the nature of the brain. Cognitive science, in its weakest form, is just a collection of fields: artificial intelligence, psychology, neuroscience, and philosophy. When there is a theoretical and empirical merging of conclusions about the features of the mind, multidisciplinary study becomes even more exciting. For instance, psychology and artificial intelligence can be merged into statistical models of how individuals act in experiments.

2.1.3 Facial expressions

Since Darwin (1872), the main concern of researchers interested in the face has been the relationship between facial movements and emotional states. Proponents of this opinion, the "Emotion View" like Eckman and Rosenberg (1997) [16], are not uniform in all respects, but believe that emotions are fundamental in explaining facial movements. Conversely, the "Behavioral Ecology View" derives from the narratives of the evolution of signaling behavior and does not interpret facial displays as expressions of emotion, but as social signs of intention that are meaningful only in social contexts. Recently, facial expressions have also been admitted as an emotional activator, as opposed to being seen only as a response to emotional.

Figure 2.2: Frames from the AffWild database which show subjects in different emotional states [17]

Emotion View

Emotion View offers a small set of basic emotions. They are reflexes, or "affecting programs," that are distinguished by natural selection, stimulated by stimuli, and accompanied by the display of facial prototypes. The six basic symbols identified by the six corresponding global terms and referred to by the following language labels are provided by Ekman and Friesen: surprise, fear, anger, disgust, sadness, and happiness [18]. Contempt has also recently become a global term. Changes in the prototype's emotional expressions have been described as reflecting the extraction of a combination of basic emotions or the effect of culture-specific conventions.

For example, there seem to be emotions associated with each of the six basic emotions mentioned above. The surprising group consists of a set of likely expressions matching to different situations: vague surprise, questioner surprise, slight surprise, moderate surprise, etc. In addition, different emotions can be placed in a single face to express sad-angry, scary and angry-afraid expressions.

Behavioral Ecology View

Facial expressions can also be considered as a communication method, the face is an independent way for transmitting communication signals. Although the human face is capable of 250K expressions, less than a hundred sets of phrases constitute unique and exact symbols. When viewed as communication signals, three main sections of signals are identified from the face.

- *Syntactic representations*: Used to press words or paragraphs.
- *Speaker displays*: Display transferred ideas.
- *Listener Comment Displays*: Used in reply to speech.

From a Behavioral Ecology View, point of view, there are no basic emotions and no basic expression. This view does not address facial representations as to the "expression" of discrete, physiological emotional states, nor as to the result of affected programs. Facial displays are supposed "significations of intent", which evolve in response to specific range pressures. The actual form of the face may depend on the speaker's personality characteristics (dominant or non-dominant) and the context (defense of the territory, access to women, restoration of borrowed property).

2.2 Machine Vision

Machine vision is a branch of knowledge that seeks to reconstruct and interpret the three-dimensional world around us by processing two-dimensional images. Simply put, machine vision means that computers can see the world with the help of cameras, understand, and even surpass human vision. [19]

Machine vision can be explored from both scientific and technological perspectives. As a science, machine vision develops the theory of intelligent systems that extract information from images, and as a discipline of technology seeks to use theories and models developed to build machine vision systems. For example, manufacturers in various industries use machine vision systems for visual inspection, which requires a high speed, magnification, 24-hour operation, and repeatability.

2.2.1 Basic concepts of machine vision

Machine vision can be considered as an interdisciplinary discipline of various sciences, so that it can be used in sciences such as computer, electricity and electronics, industry, mechanics or medicine. On the other hand, machine vision with concepts such as image processing or Video processing is closely related, so that in many cases it is not possible to draw a clear red line between them.

When we go to the basic concepts of image processing and machine vision, we come across these words Computer Vision, Machine Vision and Image Processing. "Image processing" is a comprehensive concept. By definition, it is one of the modern and diverse branches of artificial intelligence that by combining special methods and algorithms on an image, you can do different projects with specific applications. When you want to use these processing algorithms, you must use a computer, and you also have to use a camera to take a picture and send it to a computer. After the images were sent from the camera to the computer; You must use software related to this field. In this case, when you do a

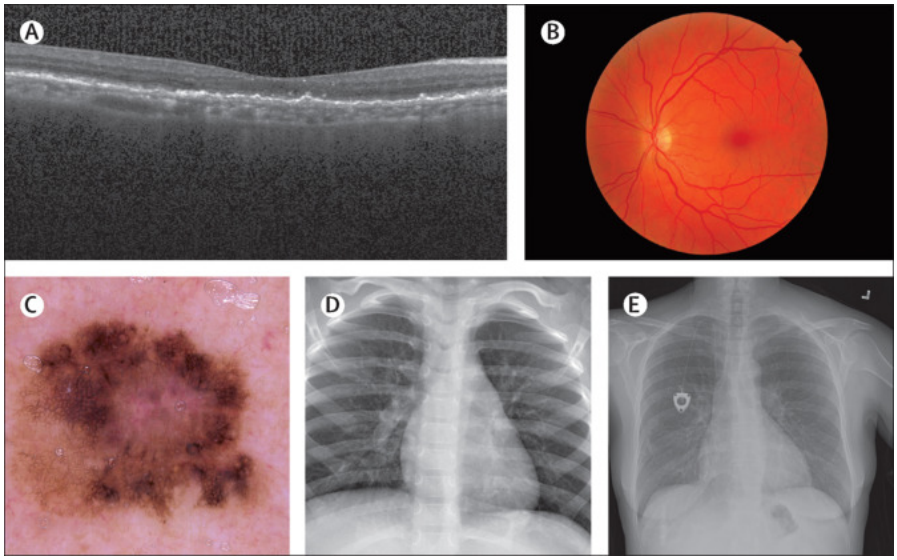

Figure 2.3: A sample of machine vision adoption in prediction of diseases. (A) Case of drusen. (B) Presence of diabetic retinopathy. (C) A melanoma. (D) Pneumonia. (E) A pleural effusion [20]

project this way; In fact, you are using a computer vision system (Computer Vision).

2.2.2 Different Techniques

Machine vision processes are categorized into three techniques:

Low Level Vision: In this technique, image processing is performed to extract a feature (edge, corner, or light stream).

Mid-Level Vision: This technique is performed using features extracted from low-level vision, object detection, motion analysis, and three-dimensional reconstruction.

High Level Vision: This technique is responsible for interpreting the information provided by mid-level vision. These interpretations may include conceptual descriptions of the scene, such as activity, intent, and behavior. This level also specifies what the lower and middle level vision should do.

2.3 Object Recognition in Machine Vision

In machine vision systems and related applications, object recognition is the process of identifying an object and its position within an image or scene. From the beginning of the field of digital image processing, machine vision, and computer vision, the act of recognition, usually using a reference data set and two-dimensional image data.

In recognizing objects in two-dimensional digital images, researchers often face situations such as "partial occlusion", "partial light conditions", and "cluttered backgrounds" in digital images.

If 3D scenes are used, some of the problems mentioned, such as partial overlap, will be eliminated in the image analysis phase (because this problem is resolved in the stereo correspondence phase). Also, in anomalous and cluttered scenes, the analysis of three-dimensional scenes may only be problematic if the system is recognizing the scene or clustering the set of three-dimensional pixels (Voxels) in order to identify the objects in the image.

Recognition of objects in machine vision systems is usually done in two stages: "Acquisition" or "Training" and "Recognition". In the first step, the machine vision system must collect the objects in the image in its memory so that in the next steps it can detect and recognize them.

The conventional method for doing this is to first create a database consisting of object models, which includes a collection of images at different angles and different lighting conditions (Lighting Condition); Various studies have shown that using such an approach is not cost-effective in terms of memory. Another method used for three-dimensional modeling of objects is "Constructive Solid Geometry" (CSG). Research has shown that using CSG methods to model three-dimensional objects is a more efficient (in terms of memory) method for storing three-dimensional information related to an object.

In the CSG method, a combination of basic graphical elements (Solid Primitives) with operations such as "Intersection", "Union" and "Set Difference" is used to form an object. Objects obtained through the CSG method can be logically represented by a "Binary Tree"; In this view, tree leaf nodes are represented by basic graphic elements and parent nodes are represented by operations defined between them (sharing, community, and set difference).

The "wireframe" model is also used to display the outline of a three-dimensional

model. However, the big problem with the wireframe model is that models produced from the same object (in different perspectives) have a similar "Projective Appearance"; As a result, this method is not suitable for three-dimensional modeling of different objects. In addition, storage methods based on three-dimensional pixels can use "Spatial Occupancy" techniques for three-dimensional modeling; In this technique, three-dimensional pixels are enclosed in a three-dimensional, discrete binary array.

In spatial occupation techniques, the object is constructed volumetrically by combining the structural information contained in binary arrays. A general version of the Constructive Solid Geometry (CSG) method is implemented in a system called ACRONYM, which stands for "A Cone Representation of Objects Not Yet Modeled", to model three-dimensional objects.

The ACRONYM system uses a concept called "swept volumes" to model three-dimensional objects. Objects (volumetric) such as "Cylinder", "Cube", "Pyramid", even a "Bottle", are all swept volumes. ACRONYM works by first sweeping objects by this system and then by clustering them, a three-dimensional model of the object is created. However, the ACRONYM system had difficulty producing complex 3D objects, and as a result, its development was halted by researchers and programmers.[21]

2.3.1 Object Recognition

Object recognition technologies can be defined in the form of computer technologies and systems, which are a set of automation tasks related to the fields of computer vision, machine vision, and image processing. In other words, the objects detection or recognition are a subset of computer technologies and systems that operate in the field of computer vision and image processing.

"Object Recognition" is one of the techniques defined in the field of machine vision and computer vision that are used to identify objects in images or videos. Image processing techniques, machine learning and "deep learning" are the best tools available for object recognition.

When people look at an image or watch a video, they can easily identify the people, objects, scenes, and "visual details" in them. The goal of object recognition systems in machine vision and computer vision is to enable the computer system to learn what is very simple for humans to do and to be able to understand the content of images and videos.

2.3.2 The Difference Between Object Detection and Object Recognition

Object detection and object recognition are similar techniques for identifying objects in images and videos, except that they work differently to identify objects. In object detection systems, the goal of the system is to find examples of objects in images. When deep learning methods are used to identify objects, object detection systems are a subset of object recognition methods; Because these methods, in addition to being able to identify objects, can also identify their position in the image. This allows the system to identify several objects in an image and locate them.

Different approaches can be used to identify objects. In recent years, techniques such as machine learning techniques and deep learning have become the most popular techniques developed to solve object recognition problems. Both techniques attempt to identify objects in images and video files. However, they work differently to identify objects. The following figure shows how machine learning methods work and deep learning in object recognition.

2.3.3 Object Detection Methods

The most important methods of object detection usually use machine learning-based approaches or deep learning-based models to identify objects in the image. In machine learning-based approaches, the properties associated with the objects in the image are first extracted using special methods. The most important feature extraction methods in machine learning-based approaches are:

- Viola–Jones object detection method based on Haar properties
- Scale-Invariant Feature Transform (SIFT)
- Histogram of Oriented Gradients (HOG)

In machine learning-based approaches, after extracting features from images or video frames, one machine learning model, such as the "Support Vector Machine", is used to identify objects in the image. Then, their position is specified in the image, and finally, the identified objects are categorized into predefined classes.

In deep-learning models, end-to-end Object Detection is provided for developers of machine vision systems and computer vision. The importance of such methods is that these systems are able to identify objects in the image without explicitly defining the properties associated with each of the defined classes. These models are usually based on "convolutional neural networks". The most important systems based on deep learning to recognize objects are:

- Region Proposals methods (to identify areas containing objects in the image) such as R-CNN, Fast R-CNN and Faster R-CNN.
- Single Shot MultiBox Detector or SSD method.
- Famous and very popular methods You Only Look Once or YOLO.

Chapter 3

CLOUD COMPUTING INFRASTRUCTURE

3.1 Cloud Environment

Cloud computing is used to provide on-demand computing services, from "applications" to "storage space" and "computing power", through the Internet with pay-as-you-go pricing.

Without having their own "infrastructure" or "data centers", companies can access everything from applications to renting storage space from a "cloud service provider". One of the benefits of cloud computing is that companies can avoid the costs and increasing complexity of owning and maintaining their IT infrastructure and instead pay for what they use and when they use it. On the other hand, cloud service providers can also reap the significant economic benefits provided by providing a similar service to a wide range of customers.

Cloud computing services offer a wide range of choices from storage, networking, and processing power to natural language processing (NLP) and artificial intelligence (AI), as well as office applications. Almost any service that does not require your physical proximity to the computer hardware you are using can now be provided through the cloud.

Cloud computing is the basis of many services. This includes consumer services such as Gmail or cloud backup of photos on your smartphone, although for services that allow large companies to host all their data and run all their applications in the cloud. Amazon also relies on cloud computing services to run its video and music streaming service (Prime Video/Music) and has several other organizations.

Cloud computing is becoming the default option for many applications: Software companies are trying to switch to a subscription model rather than stand-alone products, offering most of their applications as services over the Internet. However, there is a potential problem in cloud computing, in that it can create new risks and new costs for companies that use it.

3.1.1 History

Cloud computing has been around as a term since the early 2000s, but the concept of computing as a service has been around for much longer - since the 1960s when computer offices allowed companies to lease time on a mainframe instead of That they must buy their own.

These "time-sharing" services with the advent of personal computers made it much more cost-effective to own a computer, and then in turn increased the corporate data centers in which companies store large amounts of data.

But the concept of leasing access to computing power has reappeared repeatedly - in application service providers and network computing in the early 2000s, followed by cloud computing, which really emerged with the advent of Software as a Services (SaaS) and providers have grown to scale, such as *Amazon*.

3.2 Docker Container

3.2.1 The Reason Behind Docker

Nowadays, teams need to publish programs quickly to attract and save companies. This requirement requires software support teams to always look for solutions to save time and reduce costs. An ideal solution reduces the time spent creating and configuring deployment environments and simplifies the software deployment process.

Using software containerization technology as a solution to save time and reduce costs is a popular Idea. Independency to configure the hardware and spend time installing the operating system and software to host the deployment is one of the strengths of building containers. The containers are isolated, and several containers can work on the same machine. These abilities help using the hardware more

efficiently and can help improve the security of our application.

3.2.2 What is Docker?

"Docker is an open platform for developing, shipping, and running applications. Docker enables you to separate your applications from your infrastructure so you can deliver software quickly. With Docker, you can manage your infrastructure in the same ways you manage your applications. By taking advantage of Docker's methodologies for shipping, testing, and deploying code quickly, you can significantly reduce the delay between writing code and running it in production."[22]

3.2.3 The Docker platform

"Docker provides the ability to package and run an application in a loosely isolated environment called a container. The isolation and security allow you to run many containers simultaneously on a given host. Containers are lightweight because they do not need the extra load of a hypervisor but run directly within the host machine's kernel. This means you can run more containers on a given hardware combination than if you were using virtual machines. You can even run Docker containers within host machines that are actually virtual machines!"[22]

Docker provides the tools and platform for managing the life cycle of your containers:

- Develop the program and components that are basic blocks, using containers.
- To distributing and testing the application, the container becomes a major unit.
- Finally, place the application as a container or an orchestrated service in the production environment. Whether the production environment is a local data center, a cloud provider, or a combination of the two, it does the same thing.

3.2.4 The Docker Engine

Docker Engine lets the programmer develop, collect, send, and run programs using these components:

- *Docker Daemon*: An ongoing background process that steadily listens to Docker API requests and processes them and meanwhile manages Docker networks, storage volumes, images, and containers.
- *Docker REST API*: An API adopted by applications to communicate with the Docker daemon. Available by an HTTP client.
- *Docker CLI*: A command-line interface client to cooperate with the Docker daemon. This simplifies how to run container instances and is one of the main causes developers are interested in using Docker.

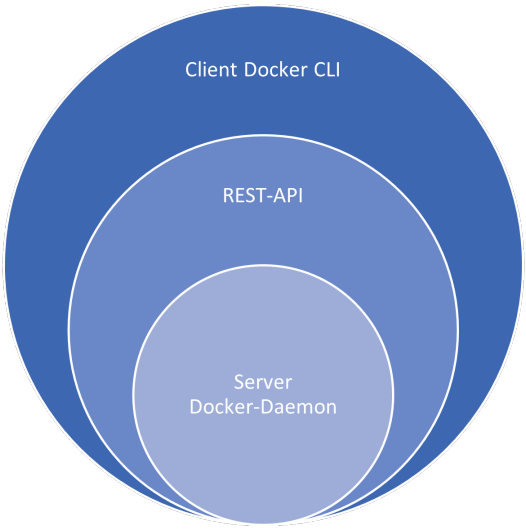

Figure 3.1: Docker Engine

3.2.5 Docker Hub

Docker Hub is a Docker container registry that is based on SaaS methodology. Docker registries are repositories that we use to store and distribute images of the container created by developers. Docker Hub is the default public registry for managing container images.

3.2.6 What is Container?

"A container is a standard unit of software that packages up code and all its dependencies, so the application runs quickly and reliably from one computing

environment to another. A Docker container image is a lightweight, standalone, executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries, and settings."[22]

3.3 Google Cloud

Google Cloud Platform is a collection of public cloud computing services provided by Google. The platform includes a range of hosting services for computing, storing, and developing applications that run on Google hardware. Software developers, cloud administrators, and other IT professionals can access Google Cloud Platform services over the public Internet or through a dedicated network connection.

Google Cloud Platform provides services for machine learning, computing, networking, storage, big data, and the Internet of Things (IoT). GCP also provides cloud security, management, and tools for developers. The main products of cloud computing in Google Cloud Platform are presented in next section.

3.3.1 Google Cloud Computing Services

The following services are provided by google cloud in order to establish different type of services.

Google Compute Engine, infrastructure as a service (IaaS), provides users with instances of virtual machines for hosting workloads.

Google App Engine, a platform as a service (PaaS) that allows software developers to access scalable Google hosting. Developers can also use the Software Development Kit (SDK) to produce software products that run on the App Engine.

Google Cloud Storage is a platform that uses as cloud storage and designed to store unstructured and large datasets. Google also gives database storage options, including Cloud Datastore for non-relational storage, Cloud SQL for fully relational storage like MySQL, and the native Google Cloud Bigtable database.

Google Container Engine is an orchestration system for Docker containers that work in the Google cloud. Google Container Engine is working based on the Google Kubernetes container orchestration engine (K8s).

Google Cloud Platform also provides services to develop and integrate applications. A real-time, managed messaging service that is known as Google Cloud Pub/Sub allows applications to exchange messages with each other. Another service is Google Cloud Endpoints that allows programmers to develop services based on the RESTful API and then make it available to JavaScript clients, Android, and Apple iOS. There are also other services provided by Google Cloud like direct network interconnection, load balancing, Anycast DNS servers, monitoring, and logging.

3.4 Establish Application on Google Cloud

3.4.1 Create Docker Container

The application that is developed is based on two separate system. One web application that is responsible for serving a HTML page and works as Client and a core application that performs classification job as Server. These two applications are connected via WebSocket to decrease the number of requests sending between Client and Server.

In order to establish this architecture over Google cloud and using the scalability feature which is one of the most important aspects of cloud environment, two containers prepared. It was also possible to install the application on two different VMs, but to make the application portable and easier to install in term of scalability, using a container was a better choice.

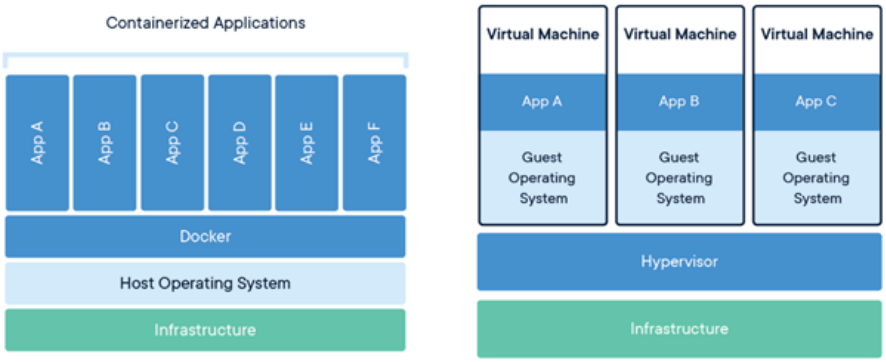

Figure 3.2: Comparison between container-based applications and traditional VM-based applications.

Here is the main advantage for choosing container-based implementation:

- *Resource efficiency*: Compared to virtualization a hardware server, separating the process level and using the container host's kernel is more efficient.
- *Portability*: All dependencies of an application are packaged in the container. This makes movement between testing, development, and production environments very easy.
- *Continuous integration and Continuous delivery (CI/CD)*: Having compatible environments and flexibility with patching has made Docker a great option for teams looking to move to the software delivery approach used by DevOps which is a modern strategy from the waterfall procedure.

3.4.2 Container-native load balancing in Google

To implement the architecture that we chose for our application we had to use the container-native load balancing. The architecture has two main section:

- External HTTP(S) Load Balancer
- Network Endpoint Group

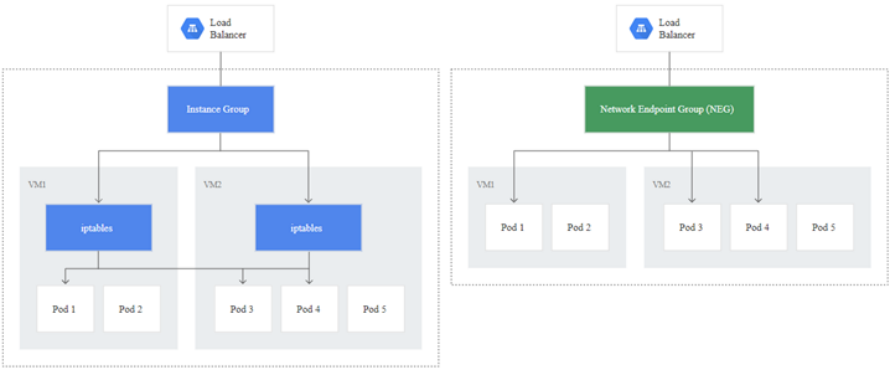

Figure 3.3: Comparison of default behavior (left) with container-native load balancer behavior.

In the figure 3.3 the different between default architecture of cloud applications and the container-native applications can be seen.

3.4.3 External HTTP(S) Load Balancing

Google Cloud HTTP (S) Load Balancing is a layer 7 load balancer that lets you run and scale your services from an external IP address worldwide. External HTTP (S) Load Balance routes HTTP and HTTPS requests to the backend that is hosted on Google Kubernetes Engine (GKE). The Kubernetes itself is provided by the Compute Engine service that is one of the Google cloud services.

This Load Balancer runs on Google Front Ends (GFE). GFEs are spread globally and work with each other using the World Wide Web and the Google Control Panel. GFEs provide cross-regional load balancing, route HTTP(S) traffic to the nearest healthy backend service with enough capacity and can deliver HTTP(S) traffic to users as fast as possible, in Premium Tier. On the other hand, in Standard Tier, load balancing is done based on region.

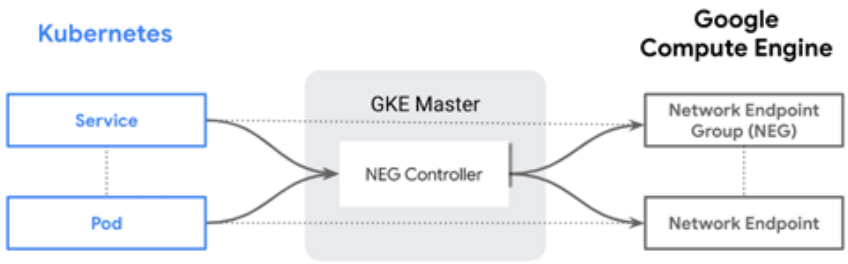

Figure 3.4: The overall scheme of communication between Kubernetes and Google Cloud Engine.

Figure 3.4 represents the communication between Kubernetes API objects and Compute Engine objects.

3.4.4 Network Endpoint Group (NEG)

"A network endpoint group (NEG) represents a group of backends served by a load balancer. NEGs are lists of IP addresses that are managed by a NEG controller and are used by Google Cloud load balancers. IP addresses in a NEG can be primary or secondary IP addresses of a VM, which means they can be Pod IPs. This enables container-native load balancing that sends traffic directly to Pods from a Google Cloud load balancer."[23]

There are two type of NEG:
Ingress with NEGs: “When NEGs are used with GKE Ingress, the Ingress controller facilitates the creation of all aspects of the L7 load balancer. This includes creating the virtual IP address, forwarding rules, health checks, firewall rules, and more. Ingress is the recommended way to use container-native load balancing as it has many features that simplify the management of NEGs. Standalone NEGs are an option if NEGs managed by Ingress do not serve your use case.”[24]

Standalone NEGs: “When NEGs are deployed with load balancers provisioned by anything other than Ingress, they are considered standalone NEGs. Standalone NEGs are deployed and managed through the NEG controller, but the forwarding rules, health checks, and other load balancing objects are deployed manually.”[24]

Since the Ingress is a recommended to implement container-native load balancing we decided to use this type of NEG. In the following figure the differences between these two scenarios have shown.

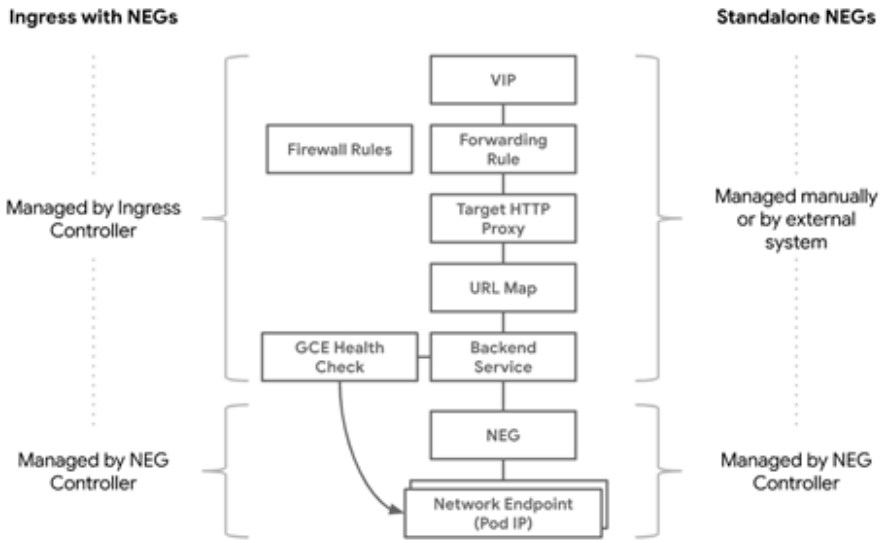

Figure 3.5: Comparison between the structure of "Standalone NEGs" and "Ingress with NEGs".

Finally, by putting different elements of this structure together, we have a container-native load balancer that provides a HTML page for clients and a core service to do classification of images.

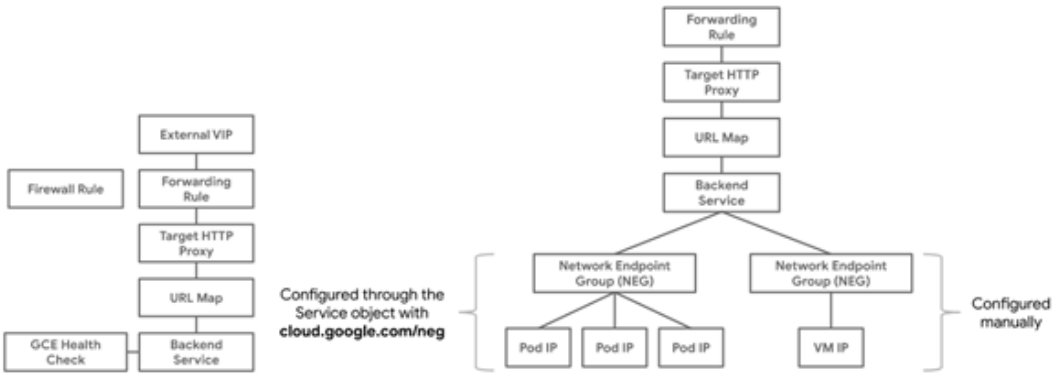

Figure 3.6: On the left side a detailed structure of the load balancer and on the right side, a complete view of the structure of the application have shown.

The figure 3.6 shows an overview of this structure.

Chapter 4

Data Collection and Analysis

4.1 Datasets

A challenging issue in computer vision is affective computing in the wild world. The current annotated facial expression databases in the wild are limited and mainly cover distinct emotions classified in 7 basic categories including happy, sad, angry, disgust, fear, surprised and neutral. The continuous dimensional model has very small annotated facial databases (e.g., valence and arousal). There are only a few datasets that provided these features in a labeled dataset.

4.1.1 Existing Databases

There are many different datasets with different characteristics available on the web. but based on needs of this research it was not possible to use many of these options. Table 4.1 represents a list of avaiable datasets.
Among all different options, many databases are created by researchers in a laboratory-controlled environment where the subjects portrayed various facial expressions, like early databases of facial expressions such as JAFFE, Cohn-Kanade, RaFD, ADFES. Also, some of these datasets have no practical information for images in terms of arousal and valence values.
Considering two main factors (bidimensional values for arousal and valence, spontaneously captured images) three major datasets can be acceptable. The following list indicates acceptable datasets:

- AfectNet
- AFEW-VA
- Af-Wild

	availably	#subjects	ethnicity	Emotion Labels / Valence / Arousal	#pic per each emotion	posed / spontaneous	video / pictures	distribution	#pics
AffectNet	No (Needs Request)	Internet	-	7 classes of basic emotions + Valence and Arousal	-	posed and spontaneous	pictures	-	1M (440k annotated manually)
Extended Cohn-Kanade Dataset (CK+)	No (Needs Request)	210 adults 18 to 50 years old 69% female	81%, Euro-American, 13% Afro-American, and 6% other groups	Anger, Contempt, Disgust, Fear, Happy, Sadness and Surprise + Action Unit Labels	-	posed and spontaneous	pictures	-	593 sequences (i.e., 7 to 60 frames) from 123 subjects. Only 327 of the 593 sequences have a given emotional class
JAFFE	yes	10 Japanese females	-	Anger, disgust, fear, happiness, sadness, surprise, and neutral	~30	Posed	picture	Normal	213
FER-2013	yes	Google search	Not mentioned	Anger, disgust, fear, happiness, sadness, surprise and neutral	Need Pre-process!	posed and spontaneous	picture	24.4% happy	28000+3500(test)
RaFD	No (Needs Request)	67 models (no glasses, make-up or facial hair)	Caucasian males, females and children + Moroccan Dutch males	Anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral	-	Posed	picture	-	More than 48000
ADFES	No (Needs Request)	22 models (10 female, 12 male)	Northern-European and Mediterranean	anger, disgust, fear, joy, sadness, and surprise, as well as contempt, pride and embarrassment.	-	Posed	648 filmed emotional expressions	-	-
CMU Multi-PIE	No (Needs Request)	337 (15 viewpoints and 19 illumination conditions)	-		-	Pose and Illumination	pictures	-	750,000 (305GB)
KDEF	yes	70 amateur actors [1]	-	Anger, disgust, fear, happiness, sadness, surprise and neutral	350 (from 5 different angles)	Posed	pictures	Normal	4900
Real-world Affective Faces Database (RAF-DB)	No (Needs Request)	Internet (labelled by about 40 annotators)	-	7 classes of basic emotions + 12 classes of compound emotions	-	posed and spontaneous	pictures	-	29672

Table 4.1: List of available datasets

4.1.2 Data Distribution

Observing the distribution and the number of samples in each section of these datasets, the AfectNet[17] chose to be used to train the CNN model explained in the next section.The figure 4.1 shows the distribution of these three datasets.

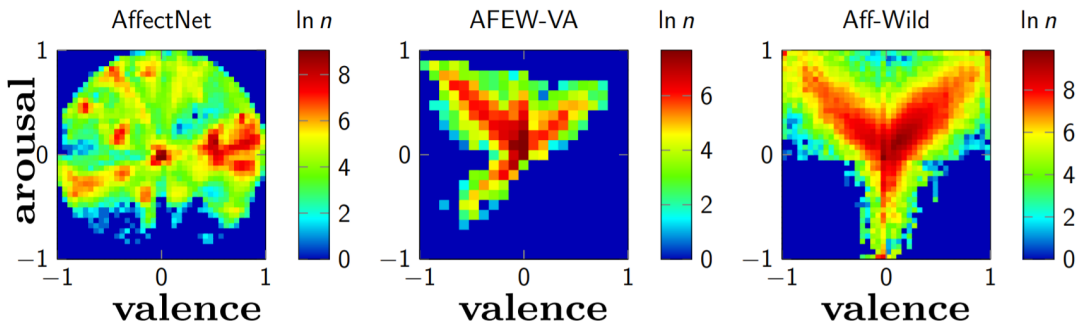

Figure 4.1: Valence/arousal ground truth distribution of the *a* AfectNet *b* AFEW-VA and *c* Af-Wild dataset

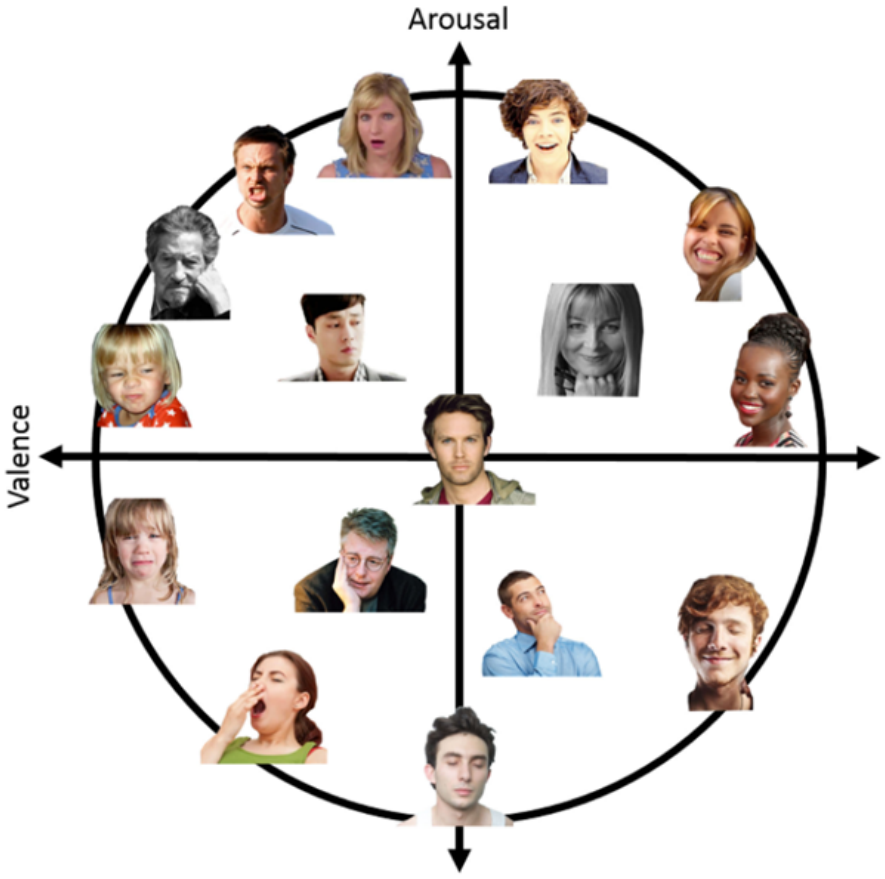

Figure 4.2: Data Distribution of AffectNet Dataset[17]

4.2 Model Architecture

The CNN model structure adopted in this thesis includes tree sub sections including:

- Face Detection by using CV2
- Feature Extraction by using CNN (VGG16)
- Prediction of Arousal/Valence by using Custom CNN

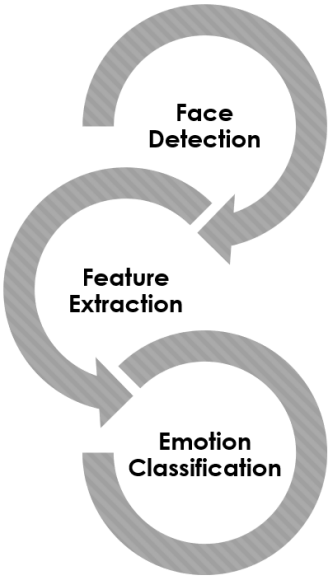

Figure 4.3: General view of model architecture

The details of the model can be seen in the figure 4.4

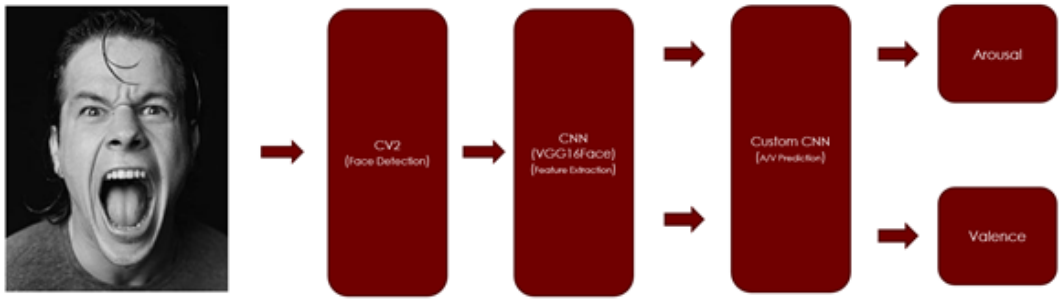

Figure 4.4: General view of model architecture

4.2.1 Face Detection

"OpenCV (Open-Source Computer Vision Library) is an open-source computer vision and machine learning software library. OpenCV was built to provide a common infrastructure for computer vision applications and to accelerate the use of machine perception in the commercial products."[25]

Figure 4.5: Face detection by using OpenCV

There are many different algorithms that can be used to identify faces, objects, classify human actions, track objects that are moving, build three-dimensional models of objects. This library also can be used to stitch images together to produce an image of a big scene with higher quality, find similar images from an image database, etc.

4.2.2 Feature Extraction

The VGG16 is one of the great architectures of the visual model to date. The most different thing about the VGG16 is that this architecture different from other architectures having many parameters, the focus is on having a 3x3 filter in convolution layers with stride equal to 1 and always using the same maxpool and padding 2x2 filter with stride equal to 2. This combination of convolution layer and maxpool layer continuously repeats throughout the architecture. Finally, it has 2 FCs (fully connected layers) followed by a softmax for the output. But in this thesis, since we are using this network as a facial feature extractor, the last two fully connected layers have been removed and the output of this model as a 1x512 NumPy array is the input for the next model. Figure 4.6 shows the general view of VGG16.

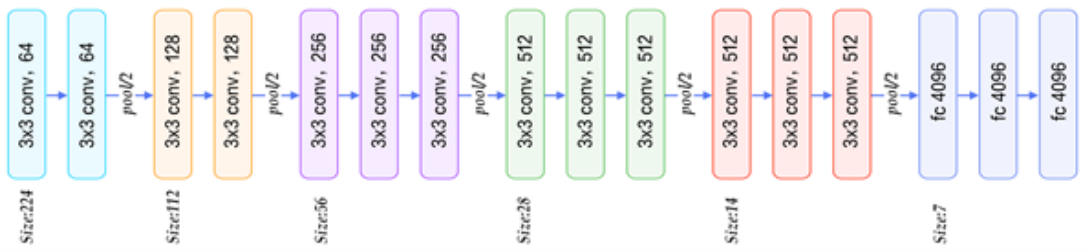

Figure 4.6: VGG16 model structure

On the other hand, Greco et al. [26] estimate the ethnic performance assessment of the ResNet-50, VGG16, and Vgg Face models trained using various training suites. The preliminary protocol includes assessments in different combinations of tests, to evaluate the generalized abilities obtained by the trained network with a specific set. Experiments show that VGG16 alongside VGG Face is very accurate in facial features extraction.

To increase the accuracy of these model by using the benefits of transfer learning, the VGG16 model that is used in this research is pretrained on “ImageNet” dataset that contains more than 14 million images.

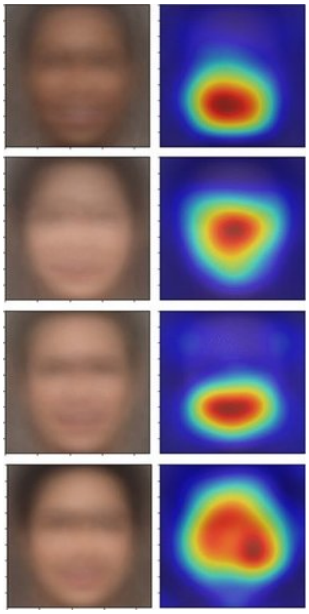

Figure 4.7: Average face images and class activation maps obtained by applying VGG16

4.2.3 Prediction

The prediction model is a combination of three fully connected layers. This combination was selected after testing different models with different complexity. Since the prediction of arousal and valence in two-dimensional space can be very complicated the model should be complex enough while the execution time of the whole structure should be minimal. Considering all the factors related to accuracy and execution time led the experiment toward a prediction model that can be seen in figure 4.8. The result of less complex model with only two fully connected layer can be seen in the figure 4.11.

The loss function that is used in training phase in this model is mean squared error (MSE). The following equation represents the loss function. Imagining Y is the correct output that the network should return (a.k.a. Target), and $\hat{Y}$ be the output that the network returns. The loss-function that is averaged over all examples in the training dataset will be :

$$\mathbb{E}_D[(Y - \hat{Y})^2] \tag{4.1}$$

The model is trying to adjust the probability distribution of $\hat{Y}$ such that it equals the probability distribution of Y. Indeed, the model is trying to make $Y = \hat{Y}$, such

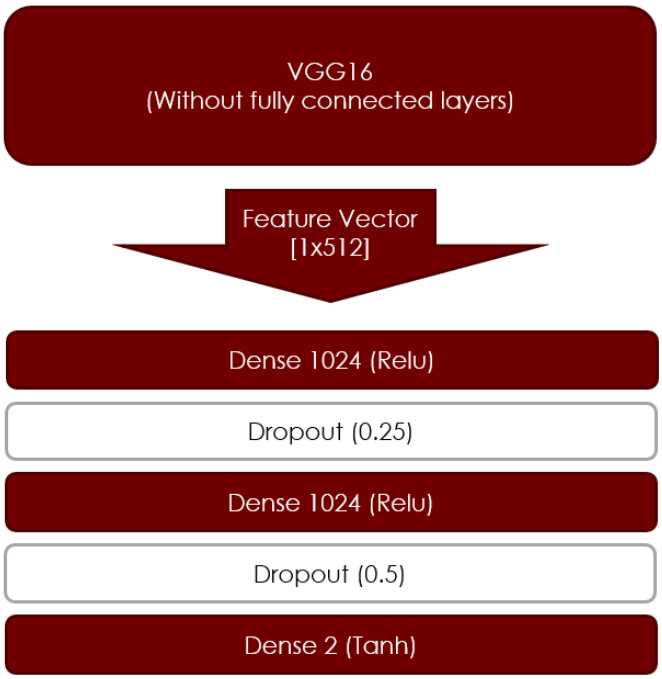

Figure 4.8: The structure of the model that is responsible to predict Arousal and Valance values

that the mean squared error becomes 0. The following equation shows the lowest possible value of the mean squared error:

$$\mathbb{E}_D[(Y - \hat{Y})^2] \geq 0 \tag{4.2}$$

Knowing that the variance of Y is always positive, then the mean loss-function has the following lower-bounds:

$$\mathbb{E}_D[(Y - \hat{Y})^2] \geq Var(Y) \geq 0 \tag{4.3}$$

4.3 Experiment

In the first step of this experiment, OpenCV is used to detect faces from each frame captured by the webcam. Since the OpenCV model works with gray scale pictures, before feeding the frames to the model, images with RGB mode converted to gray scale image with one channel. OpenCV will be loaded by an XML file that

contains weight from a pre-trained model on the "ImageNet" dataset. The output of this model is a gray scale cropped image that includes only the face of patients. The figure 4.9 shows the process of pre-processing input frames.
The output of OpenCV model will be resized to 3x224x224 to make it suitable as

Figure 4.9: The process of converting input image from the webcam to a cropped version of image that contains only the face

an input for the VGG16 model. In order to create this input, the image returned by OpenCV which is gray scale will convert to RGB mode again. VGG16 model extracts the features in form of a vector with size 1x512. Then the feature vector will be fed to the Custom CNN model with three fully connected layers. The figure 4.7 shows the average face images that is extracted by applying VGG16 on input pictures.

The results of loss and accuracy in different epochs are shown in the following figure 4.10. As the number of epoch increases to 200, the loss decreases significantly. On the contrary, the accuracy grows rapidly by increasing the number of epochs. Therefore, the best accuracy at the end is 0.695 and 0.682, in training and validation, respectively.
As it can be seen in figure 4.11 changing the number of layers in the prediction model and decreasing dropout in the last layers has a strong effect on the learning process.
In some cases when the number of fully connected layers increases/decreases too much in the model, dropout is low or the learning rate is too big, after a few epochs the model decides to stop learning and starts predicting random values which are more accurate than predicting values in the wrong way.

To optimize hyperparameters that are used to train the model, the training process performed with different Learning rates (0.0001, 0.001, 0.01), Optimizers (Adam, SGD), Activation functions (Relu, elu) and Dropouts (0.25, 0.5). Also, in the last layer two different activation function including Softmax and Tanh are used. In back propagation function, Mean Squared Error (MSE) and Mean

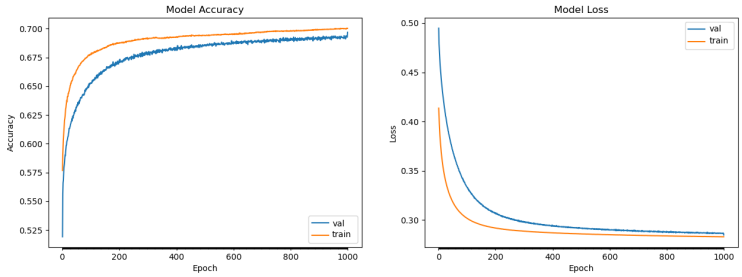

Figure 4.10: The accuracy (left) and loss (right) on our custom CNN model

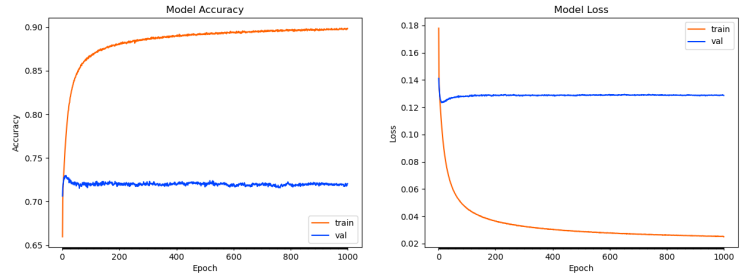

Figure 4.11: The accuracy (left) and loss (right) on our custom CNN model with less complexity in fully connected layers

Absolute Error (MAE) selected as a loss function, but since MAE is not punishing large error in prediction, the final loss function that used by the model was MSE.

Chapter 5

Conclusion

The basic emotion program presented in this system is based on the results provided by the machine learning techniques (which can identify the user's emotions with 66% accuracy in average). This system is designed to examine the emotions of patients who cannot express their emotions well, so it is used by images taken live from a smart device. Thus, the system led to the use of the Progressive Web App, which can run on any operating system and device with a modern browser and camera, where just taking a picture is enough to identify current emotions in terms of Arousal and Valence using machine learning technique.

The most problematic issue encountered in this thesis was using the webcam as an input source for application that faced the application with many problems like, lack of good illumination, low quality of images, images with the background that is hard to identify, etc.

On the other hand, finding a good dataset with well-describing labels was another issue. The existing datasets we were able to find had two problems. First, the dataset was not large enough in some area, for instance, in the area with low arousal and low valence and in the area with low arousal and high valence, the number of images is very low, meaning that there really could not be enough combinations for the model to learn from any possible conditions for new photos. Second, the values for arousal and valence annotated manually for each image were not accurate in some cases. Due to the presence of these outliers, the total accuracy dropped slightly.

Another big problem with this dataset was that the images included were not very diverse. This is mostly shot from a specific area, so if the system is used in another environment where images may have different facial features than the system, it may not work well, as the model is not used for different types of faces. Depending on the result, the characteristics can be considered as a kind of bias.

This problem can be solved if a good investment is made in building a high-quality data set from real situation in clinic with real patients, in the future. It takes a lot of effort, but it will be very useful for systems that use this data set. As the created data set may be added on top of the existing set, it is constantly added to the model accuracy and variation in detection.

It is worth noting that another possible way to extend the system is to detect emotions using an audio signal separately and to use both image and audio signal to give a combined response to detect emotions. In the web application code, there is a work code that records the sound, along with a user-friendly interface for future work around that area. Because there are external hardware devices that can improve signal quality by eliminating noise, it can be an interesting area of research, as its back and front can be enhanced by voice recognition as an additional feature. Because these parts of the project are completely modular.

As a result, the system can quickly and accurately identify and display the user's feelings by capturing the user's face using the smartphone's camera, with the advantage that it can be accessed through any app on any device. It is possible that this system will be useful for health care, hospital centers and customer service, because it can automatically detect emotions instead of asking the user accurately or decoding emotions using long tests automatically.

However, the performance and accuracy of emotions depending on the data set and models, are the main predictor variables. Therefore, the better the data set found or created in time, the better the system performance. Given that the system is modular so that it can easily replace the final model and the system works very well, according to the purpose of the project is a very good result.

Bibliography

[1] Philip Wang et al. «Use of Mental Health Services for Anxiety, Mood, and Substance Disorders in 17 Countries in the WHO World Mental Health Surveys». In: *Lancet* 370 (Oct. 2007), pp. 841–50. DOI: 10.1016/S0140-6736(07)61414-7 (cit. on p. 1).

[2] Jeremy Hogeveen, Geoffrey Bird, Aileen Chau, Frank Krueger, and Jordan Grafman. «Acquired alexithymia following damage to the anterior insula». In: *Neuropsychologia* 82 (Jan. 2016). DOI: 10.1016/j.neuropsychologia.2016.01.021 (cit. on p. 2).

[3] Alex Sel. «Predictive codes of interoception, emotion, and the self». In: *Frontiers in psychology* 5 (Mar. 2014), p. 189. DOI: 10.3389/fpsyg.2014.00189 (cit. on p. 2).

[4] Katja Schlegel and Klaus Scherer. «Introducing a short version of the Geneva Emotion Recognition Test (GERT-S): Psychometric properties and construct validation». In: *Behavior research methods* 47 (Sept. 2015). DOI: 10.3758/s13428-015-0646-4 (cit. on p. 2).

[5] Katja Schlegel, Didier Grandjean, and Klaus Scherer. «Introducing the Geneva Emotion Recognition Test: An Example of Rasch-Based Test Development». In: *Psychological assessment* 26 (Dec. 2013). DOI: 10.1037/a0035246 (cit. on p. 2).

[6] Katja Schlegel et. al. «GERT Test». In: (2020). URL: https://www.unige.ch/cisa/emotional-competence/home/research-tools/gert/ (cit. on p. 3).

[7] Tom Mitchell. *Machine Learning*. McGraw-Hill Education, 1997 (cit. on p. 3).

[8] Michael Crawshaw. «Multi-Task Learning with Deep Neural Networks: A Survey». In: (Sept. 2020) (cit. on p. 9).

[9] Rich Caruana. «Multitask Learning». In: *Machine Learning* 28 (July 1997). DOI: 10.1023/A:1007379606734 (cit. on p. 9).

[10] L. Nadel. *Encyclopedia of Cognitive Science*. Nature Publishing Group, 2003 (cit. on p. 10).

[11] Howard Gardner. *The Mind's New Science: A History of the Cognitive Revolution*. Basic Books, 1985 (cit. on p. 11).

[12] Vincent Müller. «Margaret A. Boden, Mind as Machine: A History of Cognitive Science, 2 vols». In: *Minds and Machines* 18 (Mar. 2008), pp. 121–125. DOI: 10.1007/s11023-008-9091-9 (cit. on p. 10).

[13] «Cognetive Science». In: (2020). URL: https://plato.stanford.edu/entries/cognitive-science (cit. on p. 11).

[14] Gordon Bower and John Clapper. «Experimental methods in cognitive science». In: (Jan. 1993) (cit. on p. 12).

[15] Paul Thagard. «Theory and experiment in cognitive science». In: *Artificial Intelligence* 171 (Dec. 2007), pp. 1104–1106. DOI: 10.1016/j.artint.2007.10.006 (cit. on p. 12).

[16] Lisa Barrett, Ralph Adolphs, Stacy Marsella, Aleix Martinez, and Seth Pollak. «Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements». In: *Psychological Science in the Public Interest* 20 (July 2019), pp. 1–68. DOI: 10.1177/1529100619832930 (cit. on p. 12).

[17] A. Mollahosseini, B. Hasani, and M. H. Mahoor. «AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild». In: *IEEE Transactions on Affective Computing* 10.01 (Jan. 2019), pp. 18–31. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2017.2740923 (cit. on pp. 13, 31, 32).

[18] Friesen Wallace V. Ekman Paul. «Measuring facial movement». In: *Environmental psychology and nonverbal behavior* 1 (Sept. 1976), pp. 56–75. DOI: 10.1007/BF01115465 (cit. on p. 13).

[19] Ramesh Jain, Rangachar Kasturi, and Brian Schunck. *Machine Vision*. Jan. 1995. ISBN: 978-0-07-032018-5 (cit. on p. 14).

[20] Livia Faes et al. «Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study». In: *The Lancet Digital Health* 1 (Sept. 2019), e232–e242. DOI: 10.1016/S2589-7500(19)30108-6 (cit. on p. 15).

[21] «Object Recognition». In: (2020). URL: https://www.mathworks.com/solutions/image-video-processing/object-recognition.html (cit. on p. 17).

[22] «Docker». In: (2020). URL: https://docs.docker.com/get-started/overview/ (cit. on pp. 22, 24).

[23] «Google Cloude». In: (2020). URL: https://cloud.google.com/load-balancing/ (cit. on p. 27).

[24] «Google Cloude». In: (2020). URL: https://cloud.google.com/kubernetes-engine/docs/how-to/standalone-neg (cit. on p. 28).

[25] «OpenCV Library». In: (2020). URL: https://opencv.org/about (cit. on p. 34).

[26] Antonio Greco, Gennaro Percannella, Mario Vento, and Vincenzo Vigilante. «Benchmarking deep network architectures for ethnicity recognition using a new large face dataset». In: *Machine Vision and Applications* 31.7 (Sept. 2020), p. 67. ISSN: 1432-1769. DOI: 10.1007/s00138-020-01123-z. URL: https://doi.org/10.1007/s00138-020-01123-z (cit. on p. 35).