Leveraging Acoustic Voice Characteristics for Face Recognition during the COVID-19 Pandemic

D. N. Kamalsooriya

2021

Leveraging Acoustic Voice Characteristics for Face Recognition during the COVID-19 Pandemic

A dissertation submitted for the Degree of Master of Business Analytics

D. N. Kamalsooriya
University of Colombo School of Computing
2021

DECLARATION

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge, it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: Dakshila Nayanahari Kamalsooriya

Registration Number: 2019/BA/011

Index Number: 19880111

Signature: Date: 22/11/2022

This is to certify that this thesis is based on the work of Ms. Dakshila Nayanahari Kamalsooriya under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr. Ajantha Athukorale

Signature: Date: 23/11/2022

I would like to dedicate this thesis to:

My Mother

My Sister

My Relatives

My Friends

And all the freshers in the Data Science Field.

ACKNOWLEDGEMENTS

It is with pleasure that I express my heartfelt gratitude to Dr. Ajantha Athukorale, the University supervisor who helped and guided me throughout this endeavour.

I would like to thank my mother Mrs. Ramani Ranwalage, my sister, relatives, and friends their love, support and understanding during the years of my education.

Lastly, I offer my gratitude to all of those who assisted me in providing information and extending their support toward completing this research project successfully.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 - INTRODUCTION

1.1 Project Overview

The spread of Covid-19 pandemic has drastically changed abruptly upended the lives of people all over the world. It led to a dramatic loss of human life worldwide and has also damaged the public heath, economy, and food systems. Experts have also said that although Covid-19 pandemic will be a history one day, the virus that caused it will not and therefore we will have to learn to live with SARS-CoV-2 and its descendants (Lawton, Le Page, Vaughan and Wilson, 2021). Wearing a face mask has been an important part of the response to the Covid-19 pandemic. The use of face masks by the public has significantly increased all over the world to limit the spread of the Covid-19 virus. It has become a clothing accessory that is worn every day and everywhere. Even after the pandemic, it is likely that the people will continue to wear the face mask do protect from its virus. The Covid-19 crisis also emphasized the importance of contactless operations in contact sensitive facilities avoiding physical contact and thereby to reduce the spread of the virus. As a result, face recognition has become vital in our daily lives over the past years as a contactless and a convenient method of accurate identity verification. Although a numerous number of studies on automatic face recognition has been done in the past, a very few studies has been done to recognize face while wearing a face mask. Moreover, the past studies have revealed that wearing a face mask will not affect or change certain voice characteristics. Thus, this study aims to leverage voice acoustic characteristics to improve the face recognition when wearing a face mask.

1.2 Background

The first case of Covid-19 outbreak was reported in late December of 2019 in Wuhan, China. It spread vastly across China and rapidly outside of China (Wu, Chen, and Chan, 2020). The first case of Covid-19 outside of China was reported in Thailand in mid-January. Since then, the outbreak has now spread all over the world despite the various preventive and control measures taken. Covid-19 pandemic was declared as a Public Health Emergency of International Concern (PHEIC) by the World Health Organization in late January 2020 (Wu, Chen, and Chan, 2020). The outbreak has caused a drastic loss to the lives of humans and has damaged the world economy, food systems and public health. As at May 2021, the total number of deaths caused by COVID-19 exceeds 3.3 million while the total number of confirmed cases exceeds 167.1 million. The use of face masks by the public has significantly increased all over the world to limit the spread of the Covid-19 virus. It has become a clothing accessory that is worn every day and everywhere. Various experimental studies have suggested that wearing face masks may both protect the wearer from acquiring various infections or transmitting infections (Eikenberry et al., 2020). As it serves this dual preventive purpose regardless of the type, setting, or who wears it, it offers a double barrier against Covid-19 transmission (Abboah-Offei et al., 2021). Further, the widespread mask use is a prominent feature of the relatively successful response to the first wave Covid-19 in Taiwan (Wang, Ng and Brook, 2020). Figure 1 shows the results of a survey conducted from April 9-12, 2020, in 15 major countries (Bricker, 2022). It highlights how the use of face masks have been gradually increased since the early stages of the pandemic.

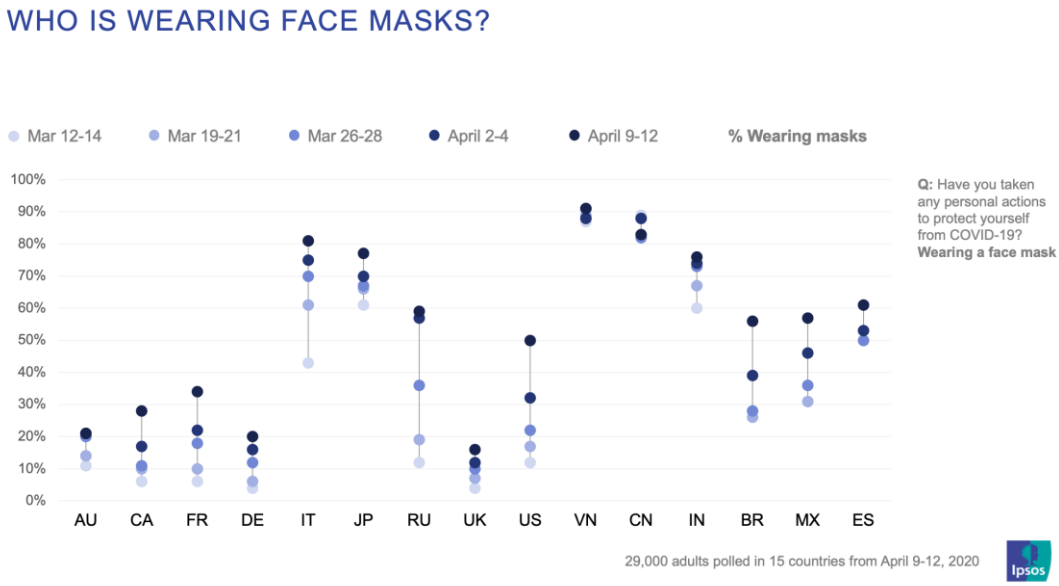

Figure 1. Who is wearing face masks? (Bricker, 2022)

The World Health Organization has recommended making wearing a mask a normal part of being around other people as a preventive measure to the spread of Covid-19. Accordingly, countries and regions across the globe have even issued policies for mask wearing so that the face mask now find much broader usage in situations where close contact of people is frequent and is inevitable, mainly inside public transport facilities, workplaces, and shops. India, Italy, Spain, South Africa, the UK, France, Vietnam are some of the countries that have mandate the use of face masks in public places whereas the USA, Canada, China are some of the countries that have recommended the use of face masks (Abboah-Offei et al., 2021). Even with the development of several vaccines, face mask wearing is still one of the most effective and an affordable way to block 80% of respiratory infections and reduce the transmission (Centers for Disease Control and Prevention, 2022).

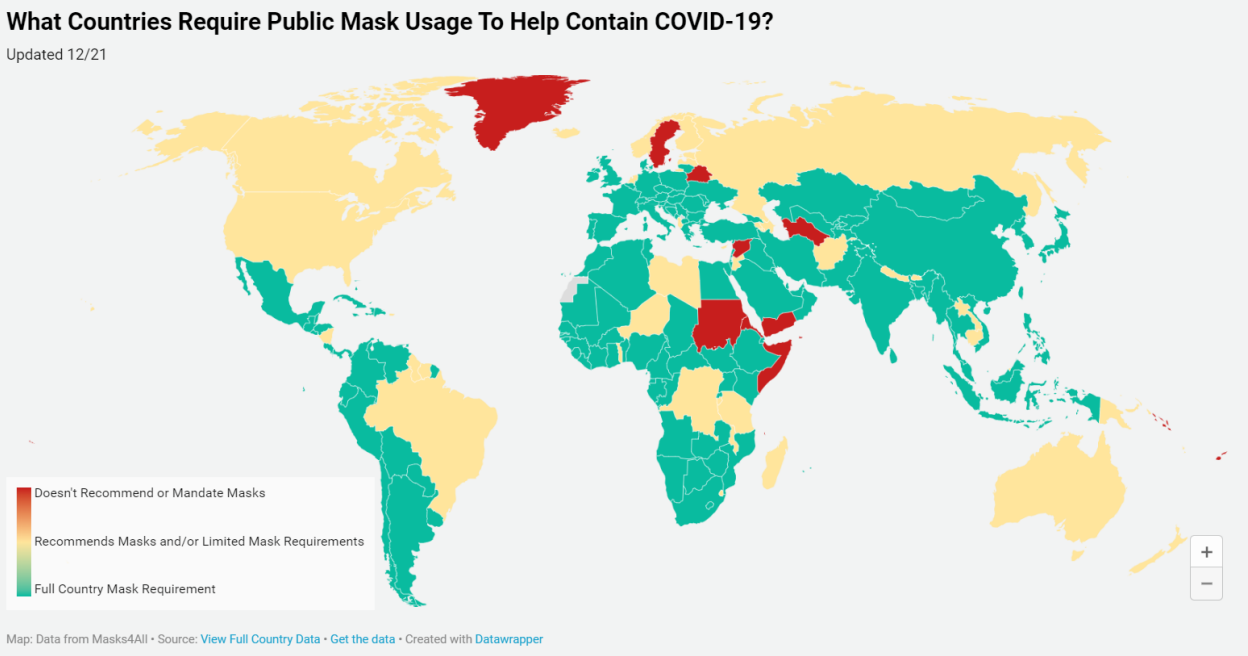

Figure 2. What countries require public mask usage to help Covid-19? (Masks4All)

Although the Covid-19 pandemic is spurring an increase in the global usage of face masks, they have been used even before that to control the transmission of various other respiratory infections. Medical masks (i.e., surgical and N95 masks) are consistently worn by healthcare workers to protect themselves from the respiratory infectious diseases as well as to protect surgical wounds from staff-generated nasal and oral bacteria (Abboah-Offei et al., 2021). Face masks are worn by individuals in some regions, e.g., in China, to avoid the exposure to air pollution.

Thus, the use of face masks has significantly increased over the last year and is probable to continue in future as well.

1.3 Motivation

The Covid-19 virus is transmitted between people both through respiratory droplets and contact routes. Studies have shown that the virus remains suspended in droplets and live on surfaces from several hours to a few days, raising concern on the hygiene and safety of sharing objects of touching surfaces (Kortli, Jridi, Al Falou and Atri, 2020). Thus, it is important to enable contactless operations in contact sensitive facilities avoiding any direct physical contact.

Over the past years, the technological advances have made face recognition technology possible in many industries. Healthcare industries make use of face recognition to patient registration. Workplaces use face recognition systems to record the attendance of their employees. It is also used in access control applications, security applications, image database investigations, general identity verifications such as in banking and surveillance (Parmar and Mehta, 2021). Thus, face recognition has become vital in our daily lives over the past years as a contactless and a convenient method of accurate identity verification.

Although face masks help to curb the spread of Covid-19 or any other infectious diseases, face masks also cover a major part of the human face affecting the performance of face recognition systems resulting in the following issues (Hariri, 2021).

1. Contactless community access control and face authentication tasks have become difficult.
2. Criminals and thieves take the advantage of mask, committing crimes and stealing without being identified.
3. Prevailing face recognition methods are not efficient when wearing face mask as the whole image of the face is in exposed.

The effect of wearing masks on face recognition is an understudied area. The National Institute of Standards and Technology (NIST) in the USA has evaluated the effect of face masks on face recognition systems provided by vendor and has concluded that the algorithm accuracy declined substantially with masked faces (Ngan, Grother and Hanaoka, 2022). Moreover, the Department of Homeland Security in the USA has conducted a similar study on a more realistic data set. It too has concluded that wearing face masks has caused a significant negative effect on the accuracy of face recognition systems [8].

Another study has been done recently compare the performance of three top-performing face recognition systems, two academic solutions and one commercial off-the-shelf (COTS) system. It

has pointed out the significant effect of wearing a mask on comparison scores separability between genuine and imposter comparisons in the investigated systems. Moreover, it points out large drop in the verification performance of the academic face recognition solutions (Damer et al., 2022).

Accordingly, it is evident that face masks have significantly challenged the existing face recognition methods. Given the current Covid-19 pandemic, it is important to improve the performance of face recognition to enable contactless and smooth-running operations and thereby help to curb the transmission of the virus from one person to another as well as to address the issues which occur to the poorly performing face recognition systems.

1.4 Objective

Objective 01: Analyze the impact of wearing a face mask to the performance of a face recognition system.

- Conduct a literature analysis of the available face recognition systems and select a face recognition system that suits the study.
- Measure the performance of the face recognition system on the collected 2/3rd of unmasked and masked facial images.
- Measure the performance of the face recognition system on the remaining 1/3rd of unmasked and masked facial images.
- Compare and analyze the results.

Objective 02: Identify if the performance of the face recognition system can be improved by making use of voice acoustic characteristics.

- Conduct a literature study on voice acoustic characteristics that would remain unchanged when a face mask is worn.
- Extract those voice acoustic features of recordings both taken without and without wearing a face mask.
- Modify the face recognition system by incorporating voice acoustic features.
- Retrain the FR system using 2/3rd of both the unmasked & masked facial images and voice acoustic characteristics.
- Measure the performance of the modified face recognition system using the remaining 1/3rd of unmasked and masked facial images.
- Analyze the results.
- Compare and evaluate the overall results.

1.5 Scope of the Study

1.5.1 In Scope

The main aim of my project is to find out if voice acoustic characteristics can be used to improve the performance of face recognition when wearing a face mask.

The data will be collected from around 25 native Sinhala speaking individuals of different age groups. Therefore, the study will be focus on the performance of face recognition in native Sinhala speakers.

An existing state-of-the-art face recognition system or an algorithm will be used to conduct the study. The performance of the selected face recognition system on the collected facial images both with and without masks will be analyzed.

Although there are various types of face masks, this study will focus on one type of a mask, mostly likely N95 mask.

1.5.2 Out of Scope

The study will not implement a new face recognition system, instead a state-of-the-art face recognition system will be used.

It will also not focus on detecting if a face mask is worn or not. Therefore, the study will be conducted or tested only on masked facial images.

1.6 Structure of the Dissertation

Project overview, Problem Statement and Motivation, Objective and Scope are discussed in **CHAPTER 1 - INTRODUCTION** while examining the Background of the problem which gives perfect edge to carry out this established problem as research. The structure of the rest of the thesis will be organized as follows.

The background and related work are described under **CHAPTER 2 – LITERATURE REVIEW**. It describes the studies already done on face recognition without face mask. Then the studies done on face occlusion and face mask detection are described. Moreover, studies done on voice acoustic analysis while wearing a face mask are also described.

CHAPTER 3 – RESEARCH DESIGN & METHODOLOGY describes the research methodology and design. It describes the phases involved in the research design and the detailed description of the first two phases will be provided in the same chapter whereas the rest of the phased will be separately described in the next chapters.

CHAPTER 4 – SOLUTION DESIGN AND APPROACH focuses on the implementation. It describes both the face recognition model implemented without using voice acoustic characteristics and with using the voice acoustic characteristics. Evaluations done to find the accuracy of the trained models implemented along with the results is described under **CHAPTER 5 – EVALUATION**.

Finally, **CHAPTER 6 – CONCLUSION & FUTURE WORK** chapter highlights the main findings and the conclusion of the project.

CHAPTER 2 – LITERATURE REVIEW

2.1 Overview

Face recognition (FR) can be considered is one of the challenging problems in the field of computer vision and image analysis. It has received a significant attention over the past years because of its application in various domains such as information security, entertainment, smart cards, law enforcement and surveillance. Occluded Face Recognition (OFR) has made face recognition even more challenging and is less covered by research. With the recent increased use of face masks Masked Face Recognition (MFR), which is a branch of OFR, and Face Mask Detection too have gained a noticeably attention. Without limiting to the facial features, the analysis of impact of face mask wearing to the voice acoustic characteristics has also received a great deal of attention over the last two years.

This Chapter presents various techniques and methods used by various researchers in Face Recognition and Voice Acoustics Analysis. Initially, the face recognition technology and its applications are discussed. Then FR history, systems, available datasets, 2D and 3D face recognition approaches, Occluded FR, Masked FR are discussed. Finally, studies conducted on analysing the impact of face mask for voice acoustic characteristics are presented.

2.2 Face Recognition (FR)

2.2.1 Introduction

Face Recognition (FR) is a technology used to recognize or identify a person's identity by analysing the pattern-based facial contours of human faces human faces (Wilmer, 2017). Over the last two decades, FR has been an important topic for both the industry and academia and has shown very rapid progress (Wilmer, 2017). It has now become one of the mostly used biometric authentication systems, given its potential in in many domains and applications such as security, surveillance, and border control. The figure below presents the typical applications of face recognition (Zhao, Chellappa, Phillips and Rosenfeld, 2003).

Areas	Specific applications
Entertainment	Video game, virtual reality, training programs
	Human-robot-interaction, human-computer-interaction
Smart cards	Drivers' licenses, entitlement programs
	Immigration, national ID, passports, voter registration
	Welfare fraud
Information security	TV Parental control, personal device logon, desktop logon
	Application security, database security, file encryption
	Intranet security, internet access, medical records
	Secure trading terminals
Law enforcement and surveillance	Advanced video surveillance, CCTV control
	Portal control, postevent analysis
	Shoplifting, suspect tracking and investigation

Figure 3. Typical applications of Face Recognition (Zhao, Chellappa, Phillips, and Rosenfeld, 2003)

2.2.2 History of Face Recognition

The most significant historical stages that have contributed to the advancement of the face recognition technology are as follows.

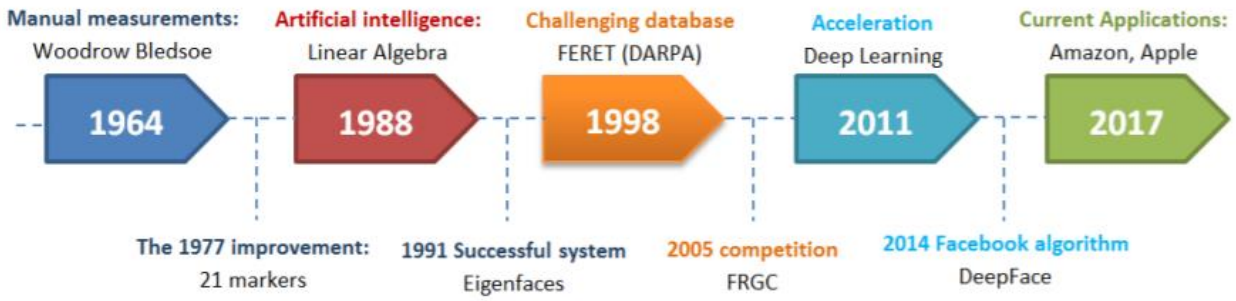

Figure 4: Primary stages in the history of face recognition (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020)

- 1964: Bledsoe et. Al studied a semi-automated face recognition computer programming. The operators were asked to enter 20 computer measures such as the size of the mouth or the eyes (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).
- 1977: 21 new features were added to the system such as the lip width and hair color (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).
- 1988: As the previous tool showed many weaknesses, Artificial intelligence was introduced to develop them. Linear algebra was used to interpret the images for efficiently and to simplify and manipulate then independent of human markers (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).
- 1991: The first successful example of facial recognition technology, Eigenfaces, which used the Principal component analysis (PCA) was presented by Alex

Pentland and Matthew Turk (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

- 1988: Face recognition technology (FERET) program was developed which provided a sizable, challenging database composed of 2,400 images of 850 persons by the Defense Research Projected Agency (DARPA) (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).
- 2005: To encourage the development of face recognition technology, the Face Recognition Grand Challenge (FRGC) was launched (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).
- 2011: Deep learning which is a machine learning method based on artificial neural networks is used. It selects the points to be compared and it performs better when more images are supplied (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).
- 2014: Deepface algorithm is used by Facebook to recognize faces which they have claimed that its method approaches the performance of the human eye near to 97% (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).
- 2017 to present: Today, facial recognition technology is used in a wide range of applications including in commercial, industrial, legal, and governmental applications (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020). E.g.,
  - Apple introduced a FR application which is now used in retail and banking.
  - Selfie Pay developed by Mastercard is a facial recognition framework for online transactions.

2.2.3 Face Recognition Systems

Three main steps are involved in developing a robust face recognition system as: Face detection, Feature extraction and Face recognition (Kortli, Jridi, Al Falou and Atri, 2020).

- **Face Detection** involves determining if the input image contains a human face or not. Many different techniques such as principal component analysis (PCA), histogram of oriented gradient (HOG) is used to detect and locate the human face (Kortli, Jridi, Al Falou and Atri, 2020).

- **Feature Extraction** involves extracting the features of the facial images detected in the previous step [4]. A face is represented with a set of features vector called ‘signature’ where the prominent facial features such as mouth, nose, eyes are described with their geometry

distribution (Kortli, Jridi, Al Falou and Atri, 2020). Eigenface, Independent component analysis (ICA), Linear discriminant analysis (LDA) are some of the techniques used to extract facial features (Kortli, Jridi, Al Falou and Atri, 2020).

- **Face Recognition** basically is about analyzing the features extracted and comparing them with the known faces stored in a specific database to identify a possible match. It involves 2 main steps as identification and verification (Kortli, Jridi, Al Falou and Atri, 2020). Verification is matching one face to another face to for an example authorize an access to an identity whereas identification compares a face to several other faces to find the face's identity. Convolutional neural networks (CNN) which is a deep learning approach, Correlation filters (CFs), k-nearest neighbor (K-NN) are some of the algorithms known to address effectively address the identification task (Kortli, Jridi, Al Falou and Atri, 2020).

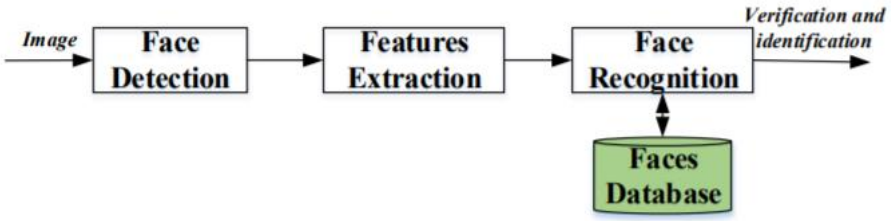

Figure 5. Face recognition structure (Kortli, Jridi, Al Falou and Atri, 2020).

In some situations, some of these steps are not separated. For an example, the facial features used for feature extraction are also frequently used in face detection. Thus, face detection and feature extraction can be performed simultaneously.

Even though a facial recognition system requires performing the three steps mentioned above, each step is considered as a critical research area as the techniques used in each step need to be improved and as they are essential in several applications. E.g., face detection for facial monitoring, feature extraction for emotion detection.

2.2.4 Assessing Face Recognition Systems

As described in the previous section, face recognition can operate either in verification or identification mode.

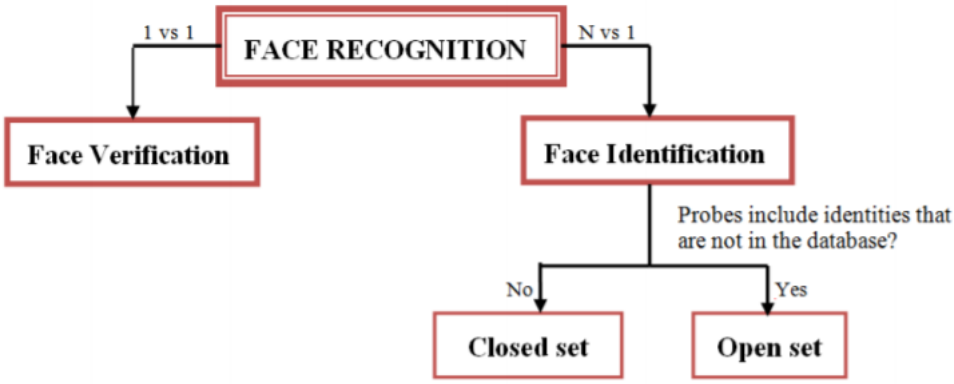

Figure 6: Categorization of FR system assessment protocols (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020)

Assessing Face Verification:

In face verification, a one-to-one comparison is performed by the system to identify if the provided face is same as another face thereby to decide whether the proclaimed identity is true or false (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020). These face recognition systems are assessed by the Receiver Operating Characteristic (ROC) and the estimated mean accuracy (ACC) (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020). True Accept Rate (TAR) and False Accept Rate (FAR) are calculated for ROC analysis as follows (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

$$FAR = \frac{False\ Positive\ (FP)}{False\ Positive\ (FP) + True\ Negative\ (TN)}$$

$$= \quad ACC = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{TP + TN + FP + FN}$$

Assessing Face Identification:

Face identification is when a face is compared to several other faces to find the face's identity. Therefore, a one-against-all comparison is done to determine the individual, without providing a prior declaration of identity. As shown in the diagram above, two test protocols may be used as open-set and closed-set (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

An open set is when the training set does not include the test identities. To measure the accuracy of an open-set model, different metrics such as the False Negative Identification Rate (FNIR) and the

False Positive Identification Rate (FPIR) are used (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

- FNIR calculates the ratio of wrongly classified cases as false even if they are true.
- FPIR calculates the ratio of wrongly classified cases as true even if they are false.

A closed set is when images from the same identities are used for both training and testing. To measure the accuracy of a closed-set model, Rank-N performance metric is used. It will return the valid user identifier within the N-Top matches (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

2.2.6 Datasets and Protocols

To study face recognition approaches, an adequate set of images should be available and accessible to the public. The diagram below summarizes the datasets made available for 2D face recognition, which can also be freely downloaded or certified with an acceptable effort.

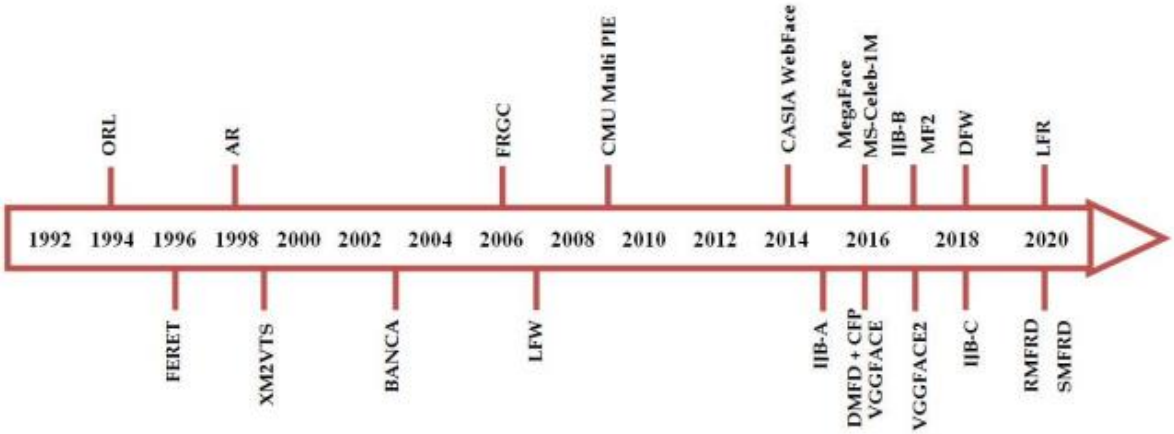

Figure 7: 2D face recognition datasets (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020)

Out of the above datasets, some of the most well-known/ recent 2D face recognition datasets are described below.

- **CASIA WebFace Dataset:** A large-scale dataset built by Yi et al. in 2014 for the face recognition task. It has around 494,414 facial images of 10,575 persons collected from the IMDb website. It is also considered as an independent training set for Labeled Faces in the Wild (LFW), which is about studying unconstrained problem of face recognition (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

- **MS-Celeb-1M:** A large scale databased released by Microsoft in 2016, consisting of around 10 million face images from 100,000 celebrities collected from the web (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

- **MF2 (MegaFace 2):** A dataset created in 2017 by Nech Shlizerman of the University of Washington, consisting of 672,000 persons and 4.7 million face images (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

- **BANCA Dataset:** BANCA Dataset was created for training and testing multi-modal biometric verification systems. Unlike the previously discussed datasets, it offers two modalities as voice and face. The voice recordings have been captured using four different European languages. Cameras and microphones in both high quality and low quality have been employed for the acquisition of data. It consists of data belonging to 2018 persons: 104 men and 104 women. The images and voice recordings have been captured in three diverse scenarios called controlled, degraded, and diverse (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

- **LFR Dataset LFR (left-front-right):** Created in 2020 by Elharrouss et al. from Qatar University to overcome pose-invariant facial recognition in the wild. To deal with the pose variation problem a CNN model for estimating pose is proposed. This CNN model is trained

Figure 8: Facial examples from BANCA database: (a) controlled, (b) degraded, (c) adverse (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

using a self-collected dataset constructed from the standard datasets - LFW, CFP, and CASIA-WebFace, consisting 3 classes of face image captures - left, front, and right side. Accordingly, images of 542 identities are gathered (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

- **RMFRD and SMFRD - Masked Face Recognition Dataset:** Unlike the previous datasets, this dataset includes three types masked face images as:
    - Masked face detection dataset (MFDD): Can be utilized to train a face mask detection model with precision.
    - Real-world masked face recognition dataset (RMFRD): Consists of images of 525 individuals both with and without masks collected from the internet.

- Simulated masked face recognition dataset (SMFRD): Alternative means to place masks on the standard large-scale face image datasets are used (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

2.2.7 2D Face Recognition Approaches

2D FR systems can be classified in to three approaches based on their detection and recognition method as shown below (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

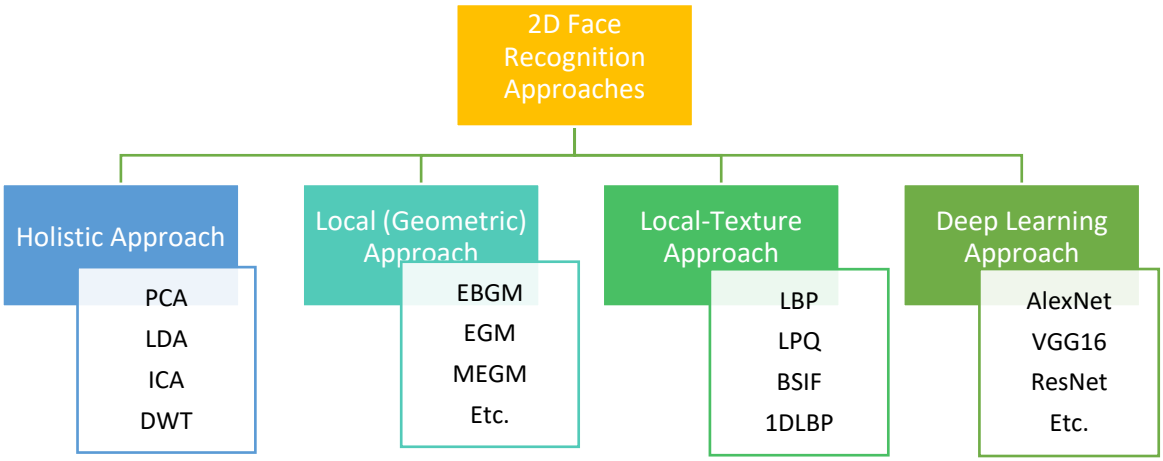

Figure 9: 2D Face Recognition Approaches

2.2.7.1 Holistic Approach

Holistic approach processes the whole face by representing the face image by a matrix of pixels. This matrix is then converted to a feature vector to facilitate their treatment. Thereafter, the feature vector is implemented in low dimensional space (Kortli, Jridi, Al Falou and Atri, 2020). Holistic approach can be divided into two categories as linear and non-linear.

Eigenfaces (Principal Component Analysis – PCA), Fisherfaces (Linear Discriminative Analysis – LDA) and Independent Component Analysis (ICA) are some of the most popular linear techniques of linear approach. Kernal PCA (KCPA), Kernel Linear Discriminant Analysis (KDA) are some of the non-linear techniques (Kortli, Jridi, Al Falou and Atri, 2020).

Holistic methods are prone to context changes and misalignments. Therefore, in majority of the cases, the face must be cut manually from the image. Further, it is required to enforce geometric consistency in all facial images as the data is viewed as a single matrix. The facial images should be carefully matched within a standard frame of reference as even a minor error in face orientation could cause considerable facial classification errors.

2.2.7.2 Local (Geometrical) Approach

Local approach treats only some facial features unlike the other approaches. The main objective of the local approach is to discover distinctive features. They are sensitive to facial expressions, occlusions and pose as they treat only some facial features (Kortli, Jridi, Al Falou and Atri, 2020). These techniques are of major two types as local appearance-based techniques and key-points-based techniques.

Local appearance-based technique is a geometric technique which will represent the face image by a set of distinctive vectors with low dimensions or small regions. The technique will the focus on critical points of the face such as nose, eyes, and mouth to extract more details. These local features will then be described through pixel orientations, histograms, geometric properties, and correlation planes. (Kortli, Jridi, Al Falou and Atri, 2020). A major drawback of these appearance-based or geometric techniques is that they involve perfectly aligned facial images. The facial images should be aligned to possess all referential features such as mouth, noise, and the eyes, displayed at the corresponding place's feature vector and therefore the facial images must be manually arranged accordingly (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

Key-points-based techniques detects specific geometric features based on the geometric information of the face surface such as the distance between the eyes, the width of the forehead etc. Key-point detection and feature extraction are the two basic steps of this technique. The first step focuses on performance of the detectors of the key-point features of the face image whereas the second step focuses on the representation of the information carried with the key-point features of the face image (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

2.2.7.3 Local-Texture Approach

Local-texture approaches focuses on the knowledge about the texture information, and they play a significant role in pattern recognition and computer vision. These techniques can be divided into two categories as statistical and structural methods. Also, they are distinctive, resilient to monotonic gray-scale changes, poor lighting, variance in brightness and does not need segmentation (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020). The local-texture descriptors have gained more attention and were introduced to many applications as well (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

2.2.7.4 Deep Learning Approach

Deep learning or deep artificial neural networks has grabbed the attention in pattern recognition tasks over the past few years (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020). It is a branch of machine learning that employs successive hidden-layers of information-processing levels, hierarchically organized for representation or pattern classification and feature learning. The three principal reasons for the prominence of deep learning as stated by Deng and Yu are:

- the dramatic growth of the processing capacity.
- drastic reduction in computer hardware costs.
- the recent progress in machine learning studies.

Deep learning is used in diverse other applications such as speech recognition, handwriting recognition, audio processing, information retrieval and has scored successful results (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

Deep learning can be categorized as follows depending on how the technique and architecture are used (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

1. **Supervised (or discriminative)** – Convolutional Neural Networks (CNN)
2. **Unsupervised (or generative)** – Recurrent Neural Network (RNN), Auto Encoder (AE), Boltzman machine (BM), Sum-product Network (SPN)
3. **Hybrid** – Deep Neural Network (DNN)

The CNN model is also commonly used for facial recognition activities because of its proven success in the field (Luo et al., n.d.; Li et al., 2014). CNN consists of filters/ kernels/ neurons with learnable parameters of weights and biases that have been added. Each of the filter will take an input, makes a convolutional and follows it with a non-linearity. The main components of a CNN are the convolutional layers, pooling, rectified linear unit, and a fully connected layer. Convolutional layer is the building block that will extract features from the input data from which a feature map will be obtained. Feature map or the activation of one layer will be fed as an input to the next layer. Pooling layer will reduce the dimensionality of the feature map without removing the crucial information. Rectified layer is a non-linear operation, which will involve units that use the rectifier. Finally, it is via the fully connected layer that the high-level reasoning in the neural network is done after applying various convolutional layers and max-pooling layers (Adjabi, Ouahabi, Benzaoui and Taleb-Ahmed, 2020).

Hybrid approach makes use of subspace and local techniques to offer better performance for face recognition (Kortli, Jridi, Al Falou and Atri, 2020).

In the last step of the face recognition system, the face identified in the input image is compared with the known faces stored in the database. Several techniques are used to do this comparison (Kortli, Jridi, Al Falou and Atri, 2020). One technique is measuring the similarity or distances of features and a few such techniques are Euclidean distance, Peak-to-correlation energy (PCE) or Peak-to-Sidelobe Ratio (PSR), Bhattacharyya distant Kortli, Jridi, Al Falou and Atri, 2020). Face classification techniques are also used such as the Support Vector Machines (SVMs), k-Nearest Neighbor (k-NN), K-means and Deep Learning (DL) (Kortli, Jridi, Al Falou and Atri, 2020).

As a result of the technological advances in face recognition, various computer vision algorithms have also been developed such as OpenCV and dlib libraries (Boyko, Basystiuk and Shakhovska, 2022). A study that has been done to compare the performance of these two libraries have concluded that OpenCV library is more productive and has better performance for face detection and recognition (Boyko, Basystiuk and Shakhovska,2022).

2.2.2 Occluded Face Recognition (OFR)

Face Occlusion refers to extraneous objects that hinder the face recognition such as face covered with a scarf, wearing glasses, cap, mask or even beard (Damer, Henry Grebe and Chen, 2020). It is a key limitation of real-world 2D face recognition. Part based method, feature based method and fractal-based method are some of the methods used to solve occlusion (Azeem, Raza, Sharif and Murtaza, 2013).

Recently, a face-eye-based multi-granularity recognition model built by applying different attention weights to the key features in visible part of the masked face, such as face contour, ocular and periocular details, forehead has been proposed to for masked face recognition (Wang et al., 2020).

2.2.3 Masked Face Recognition (MFR)

Masked face recognition (MFR) is a branch of occluded face recognition (OFR) with prior knowledge about the targeted face's occluded area. A wide range of situations and circumstances such as pandemics, laboratories, medication operations, or immoderate pollution have imposed that people wear masks in which faces are partially hidden or occluded. Recently, the COVID-19 pandemic has increased the use of masks all over the world to avoid spreading or being infected with the virus.

Wearing a mask is considered the most difficult facial occlusion challenge since it occludes a major part of the face including the nose (Vu, Nguyen and Pham, 2021). The National Institute for Standards and Technology (NIST) have presented the performance of a set of face recognition algorithms developed and tuned after the COVID-19 pandemic, following their first study on pre-COVID-19 algorithms (Alzu'bi et al., 2021). They have then concluded that many of the face recognition algorithms evaluated after the pandemic shows a performance degradation when faces are masked. Further, the performance of the recognition algorithms deteriorated when both the enrolment and verification images are masked (Alzu'bi et al., 2021).

Deep learning technologies have made great breakthroughs in both the theoretical progress and the practical application. It has further become a frontier research direction in the field of computer vision and therefore majority of the face recognition systems have moved to deep learning approaches (Alzu'bi et al., 2021).

2.3 Voice Acoustics

Voice Acoustics is an area which studies the speaking voice or the verbal communication as well as science of singing (Voice Acoustics: an introduction to the science of speech and singing, 2022). While image processing techniques have been used in many studies for face recognition, a few other studies have been done to evaluate the effect of wearing face masks on the speech or voice acoustic characteristics.

A study done very recently has compared acoustic voice measures in recordings of 16 adults producing standardized tasks with and without wearing either a surgical mask or a KN95 mask (Nguyen et al., 2021). The standardized tasks were as follows (Nguyen et al., 2021):

- 3 Repetitions of the sustained vowel – /a/ for at least 10 s.
- the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) phrases53.
- the Rainbow Passage.

These acoustic data have then been analyzed using Praat tool. The study has concluded that wearing a mask has no impact on the mean spectral level at 0–1 kHz, the cepstral peak prominence (CPP) and the vocal intensity values whereas it has an impact on the mean spectral level at 1–8 kHz region and Harmonics-to-noise ratio (HNR) values (Nguyen et al., 2021).

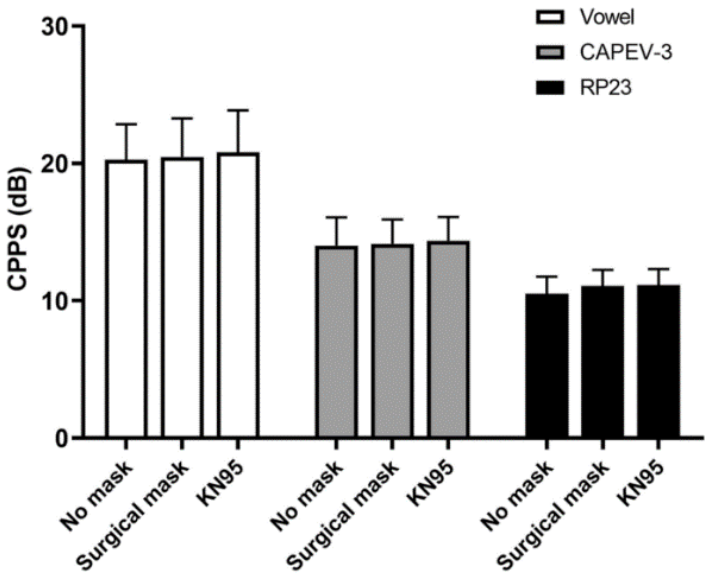

Figure 10: Cepstral peak prominence smoothed (CPPS) for all conditions (Nguyen et al., 2021)

Although face recognition is a widely studied areas, masked face recognition is an understudied area. Further, no research done on incorporating voice acoustic characteristics to improve the performance of masked face recognition can be found.

2.4 Multi-modal Biometric Verification

Multi-modal biometric refers to the use of a combination of two or more biometric modalities in a person identification of a verification system. The face recognition systems discussed in the previous section are unimodal as those will consider only the facial images. Unlike unimodal systems, multi-modal systems address the problems of non-universality, spoofing attacks, noisy data, inter-class similarities, and intra-class variations (P. S and J. B, 2013). These are widely used in many civilian applications such as ATM security and other banking securities, check cashing and credit card transactions.

There are two main phases in the multi-modal system as the enrolment phase and the authentication phase. In the enrolment phase, the biometric measures of a user are captured and thereafter stored in the database as a template for that user, that will then be used in the authentication phase. In the authentication phase, traits of the user are again captured and are then used to either identify or verify a person (P. S and J. B, 2013).

The decisions can be made at various levels of fusion such as feature level fusion, matching score-level fusion and decision level fusion (P. S and J. B, 2013).

To measure the accuracy of a multi-model biometric verification system, the metric Genuine Acceptance Rate (GAR), False Rejection Rate (FRR), False Acceptance Rate (FAR) and Equal Error Rate (ERR) are used (P. S and J. B, 2013).

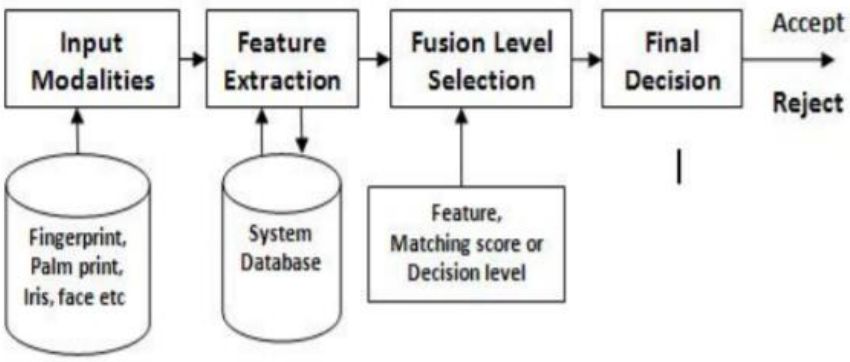

Figure 11: Block diagram of a general multimodal biometric

In literature, there are many studies done on multimodal biometric verification. Hariprasath. S et al. has given an approach using the iris and the palmprint as the modalities and using Wavelet Packet Transform [WPT] and score level fusion which has given a higher accuracy (S. Hariprasath and T. Prabakar, 2012). Another study has proposed a system using face, fingerprint, and iris as the

modalities and Bayesian decision rule-score level as the fusion technique (T. Murakami and K. Takahashi., 2011). There are also systems proposed using face, speech, and signature as the modalities where the final decision has made at matching score level with sum rule (Kartik, Vara Prasad and Mahadeva Prasanna, 2008).

However, no multimodal biometric systems have been proposed in the literature to identify masked people and to use in the COVID-19 era to enable contactless and smooth-running operations and thereby help to curb the transmission of the COVID-19 virus from one person to another.

2.5 Evaluation Metrics

Various evaluation metrics have been reported in the literature to evaluate the performance of face recognition systems. By using standard datasets, the performance has been compared with the other systems. The Categorical Crossentropy has been used as a loss function in a study where a CNN model has been used to develop a face recognition application for a biometric system (Said, Barr and Ahmed, 2022). This function is used generally for single label classification tasks, which means that each input image must belong to one output class. It gives an idea as to how incorrect the prediction of the neural network is (Said, Barr and Ahmed, 2022).

Similarly, different evaluation metrics have been used in the literature to evaluate the performance of multi-modal personal identification or biometric verifications. Out of them, accuracy is the most used evaluation metric of multimodal identification systems.

A study that has used both audio-visual feature level fusion for personal identification where both the facial and speech modalities have corrupted data with lack of prior knowledge about the corrupted data. Moreover, a limited amount of training data for each modality has been used such as, a single training facial image per person and a short training speech segment (McLaughlin, Ming and Crookes, 2022). The experiments of this study have been carried out on a bimodal data set created from the SPIDER speaker recognition database and AR face recognition database and thus, the results have been compared with the literature. The experiments have been conducted in three major parts. First, the speech modality alone, secondly, image modality alone has been considered and finally, both modalities have been considered. For each part, accuracy value has been calculated to measure the performance (McLaughlin, Ming and Crookes, 2022).

Another study that has proposed a deep multimodal fusion network to fuse three different modalities (face, iris, & fingerprint) for personal identification has used the classification metrics accuracy and Recall@K to measure the performance (Soleymani, Dabouei, Kazemi and Dawson, 2022). As they have described, the classification accuracy is the fraction of correctly classified samples out of all the classifications. The Recall@K measure is the probability that a subject class is correctly classified at least at rank – k, while the candidate classes are sorted by their similarity score to the query samples (Soleymani, Dabouei, Kazemi and Dawson, 2022). The Recall@K value is calculated per each class, and then is averaged over all available classes. Two challenging multimodal biometric databases have been used to evaluate the accuracy and Recall@K measures. Those databases are BioCop multimodal database and the BIOMDATA multimodal database.

A CNN based multimodal biometric identification system using the fusion of fingerprint, finger-vein and face images has used the publicly available SDUMLA-HMT real multimodal biometric database to evaluate the performance (Cherrat, Alaoui and Bouzahir, 2020). This study too has used accuracy as well as loss to evaluate the performance and has compared it with the accuracy of other identification systems proposed in the literature.

CHAPTER 3 – RESEARCH DESIGN & METHODOLOGY

3.1 Overview

The research design phase plays a major role in research. It defines the structure, or the systematic approach followed in to solve the problem addressed in the research and thereby giving direction and systemizing the research. In this chapter, the design phases of the research methodology that will be followed in the study is presented.

The research design and methodology of the study is divided into to five main phases as shown in the below chart excluding the problem definition and literation review phases as they are already presented in the previous chapters. The phases data collection planning, data collection and data preparation are discussed in detail in this chapter itself whereas the last two phases design & implementation and evaluation will be discussed in the next chapters in detail.

3.2 Data Collection Planning

In the literature review, it was found out that there are several datasets developed for the face recognition and masked face recognition tasks. As described in **2.2.6 Datasets and Protocols**, these datasets are freely accessible to the public and contains many images collected from many individuals. But this study requires both face images and voice recordings of the same set of individuals both with and without the face mask as the aim of the research is to study how voice acoustic features can be used to improve the masked face recognition performance. Thus, data must be collected from the scratch for this study.

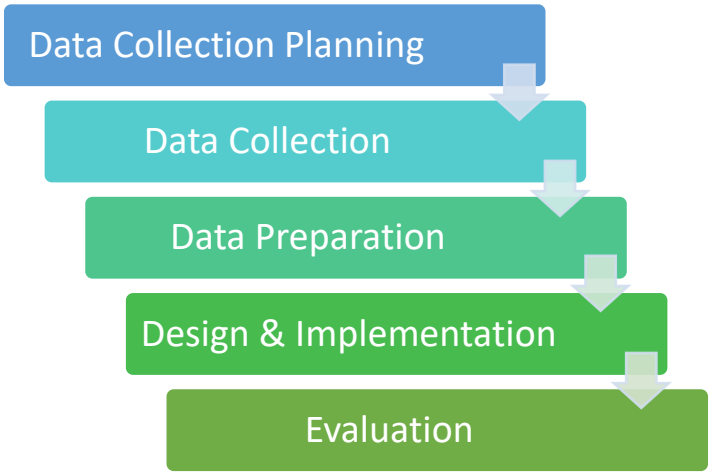

Figure 12: Main phases of the Research Design

Before starting the data collection, it is required to plan the data collection to ensure that the collected data is useful and appropriately collected. This section describes the data collection plan prepared for the study.

3.2.1 Sample Selection

Details of the target population from which the data will be collected are as follows:

Table 1: Sample Description

Age Group	12 – 15 years
Gender	Both males and females
Nationality	Sri Lankan
Mother Tongue	Sinhala

To select a sample population with the size of around 35 out of the above target population, convenient sampling method will be used. Convenient sampling is a non-probabilistic sampling method that involves the sample being drawn from that part of the population that is close to hand (Convenience sampling - Wikipedia, 2022). Considering the current situation in Sri Lanka and the Covid-19 pandemic, it is not feasible to go for a probabilistic sampling method. Therefore, convenient sampling method was selected as it is more cost effective, speedy, and easy.

3.2.2 Data Types

The data that should be collected are as follows:

- Face images both with and without wearing a mask.
- Voice recordings both with and without wearing a mask.

All these data should be collected from the same set of individuals.

3.3.3 Data Collection Settings, Standards & Tools

To ensure that the data are collected consistently and to maintain the uniformity, a set of common guidelines and settings was defined as follows.

3.3.3.1 Face Image Collection

A Canon EOS 300d camera will be used to take facial images. Participants will be instructed to wear the N95 masks to take facial images with masks. When wearing the masks, they will also be asked to use the highest level of fitting to ensure maximal barrier level and, they to press the nose metal bar so it fits tightly to the nose contour.

All the photos will be taken with the same camera setting. The photos will be taken for the default camera settings and in a white background.

3.3.3.2 Voice Recording Collection

A Sony Digital Voice Recorder ICD PX 470 will be used to record voice. Participants will be asked to read the following standardized tasks both with and without masks:

- Sustain voicing of the vowel /a/, at a comfortable pitch and loudness levels for at least 10 seconds after a deep breath.
- The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) – 3rd phrase (CAPEV-3)
- The Rainbow Passage – 3rd sentence

The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) phrases have been developed as a tool for clinical auditory-perceptual assessment of voice with the primary purpose of describing the severity of auditory-perceptual attributes of a voice problem and thereby to communicate among the clinicians (Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ASHA Special Interest Group 3, Voice and Voice Disorders, 2022). The secondary purpose is to contribute to the hypotheses regarding the anatomic and physiological based of voice problems and to examine the need of additional testing. However, it is not intended to use only in determining the nature of voice disorders but can be used as a standardized voice tasks in other areas are well (Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ASHA Special Interest Group 3, Voice and Voice Disorders, 2022).

Nguyen et al used the 3rd phrase of CAPE-V phrases in their study to assess the voice acoustic characteristics with and without using face masks (Nguyen et al., 2021). The 3rd phrase (CAPEV-3) - '*We were away a year ago'*, is an all-voiced sentence (Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ASHA Special Interest Group 3, Voice and Voice Disorders, 2022). Therefore, this phrase was selected to use as the first vocal task in this study.

The rainbow passage is a short passage that is used by speech therapists in clinics to analyze the vocals of the patients. It consists of alliteration, unusual consonants, and vowel combinations and short and extended to assess speech and breathing patters (The Rainbow Passage | Caroline Wright – Artist, England, 2022). Nguyen et al used the 2nd and 3rd sentences of the rainbow passage in their study to assess the voice acoustic characteristics with and without using face masks (Nguyen et al., 2021). Therefore, the 3rd sentence of the rainbow passage was also used in this study for the voice recording task.

Sustained voice of a vowel is another standardized task that is commonly used in voice disorder assessments but can also be used in other voice analysis tasks (Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ASHA Special Interest Group 3, Voice and Voice Disorders, 2022). As the previous two tasks, this task too has been used to assess the voice acoustic characteristics with and without wearing face masks (Nguyen et al., 2021). Therefore, sustaining a vowel for at least 10 seconds was selected as the 3rd task on voice recording.

Although the participants will be native Sinhala speakers, they will be English-literate and therefore will be able to read the selected English tasks.

3.3.4 Obtaining Informed Consent

Obtaining informed consent is one of the founding principles of research ethics. The objecting of gaining informed consent is that the participants can take part in the research freely (or voluntarily) with full information about what it means for then to take part, and that they give consent before taking part in the research. It is important that the participants understand what the research is and what they are consenting to.

As personal data will be collected in this study, it is important to obtain informed consent from the participants before collecting data. To achieve this, the following letter was prepared to distribute among the participants. The letter prepared and used for this can be referenced in **APPENDIX A - INFORMED CONSENT LETTER.**

3.3 Data Collection

A Grade 8 class of Piliyandala Central College was selected to do the data collection. First, the permission to collect data was taken from the principal of the school and the class teacher. Then the informed consent was taken from 35 students in the class. Out of the 35 students, 14 were female students and 21 were male students.

All the selected students were given N95 masks of the same colour and shape. Then, the data collection was carried out in the same setting and using the same devices by following the standards described in the previous section. First, the face images and voice recordings were captured by wearing the face masks and then the face images and voice recordings were captured without the face mask.

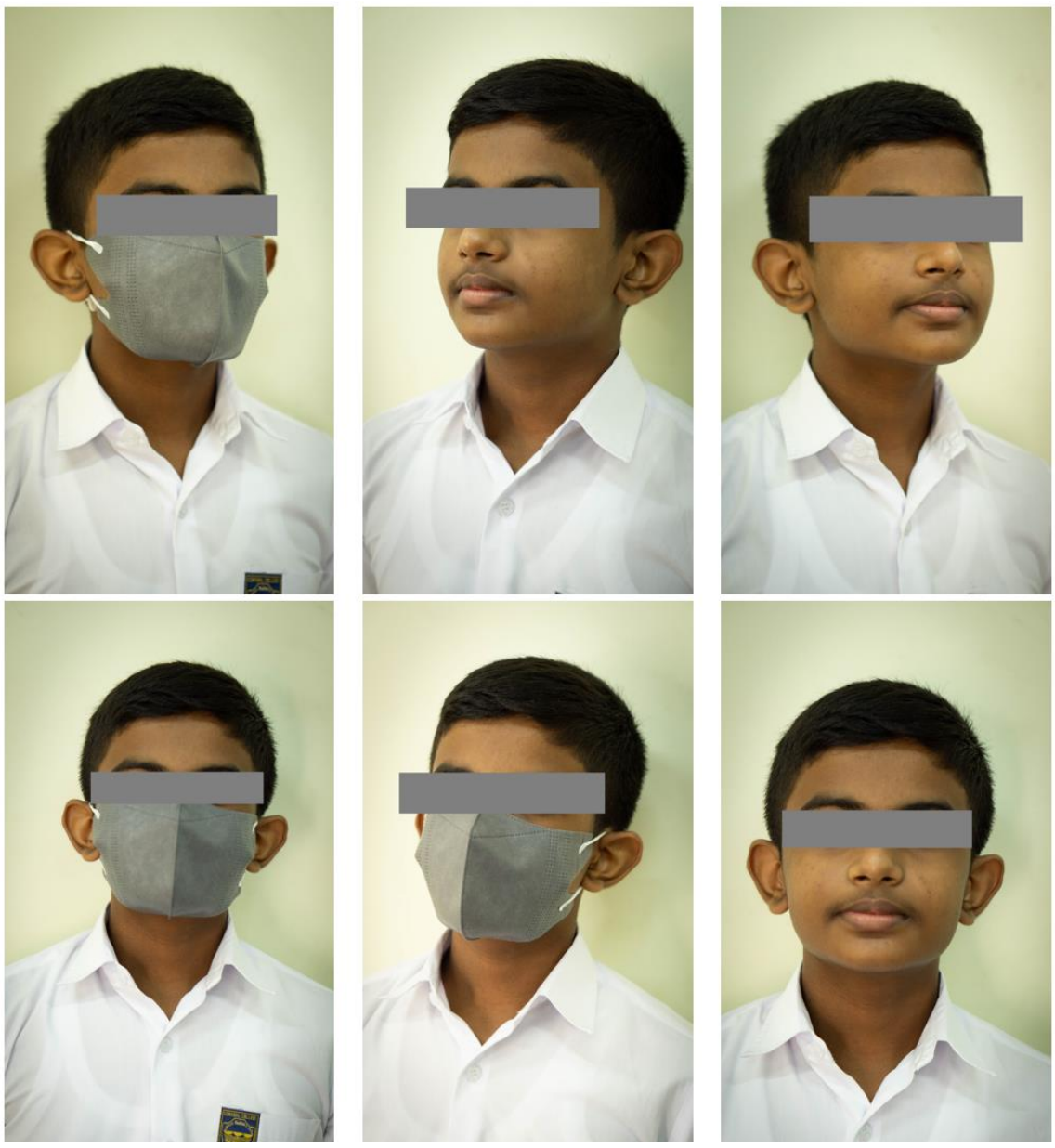

Figure 13: Face images collected from a male student

Figure 14: Face images collected from a female student

Similarly, the voice recordings too were collected in the same setting using the Sony Digital Voice Recorder ICD PX 470 for the 3 tasks listed in the previous chapter.

Accordingly, a total of 210 face images (35 participants * 6 face images with and without face masks) and 210 voice recordings (35 participants * 6 voice tasks with & without face masks) were collected. All the collected data were stored separately so that anyone else will not have access to the dataset to maintain and safeguard the privacy of the participants. A back-up of the dataset was also taken as a contingency plan in case the data gets deleted or lost.

3.4 Data Preparation

3.4.1 Preparation of Face Images

All the face images were cropped so that only the face part is kept, and the other unnecessary background is removed. Out of the 6 face images collected from a person, the following images were taken for training the model and testing the model:

- 4 images (3 without face mask & 1 with the face mask) were taken for training.
- 2 images (with the face mask) were taken for testing.

The 4 images taken from each person was put into the same folder so that there will be 35 folders each containing 4 face images. These folders were named as ‘face1’, ‘face2’, ‘face3’ etc.
All the 70 test image files (2 * 35) were put into the same folder.

3.4.2 Preparation of Voice Recordings

All the captured voice recording files were in mp3 file format. They were converted to the wav file format using a MHAudioConverter software for better processing using Tensorflow.

As it was done for the face images, the 6 voice recordings taken from each person were too divided for training the model and testing the model as follows.

- 4 voice recordings corresponding to the face images selected (3 without face mask & 1 with the face mask) were taken for training.
- 2 voice recordings corresponding to the face images selected (with the face mask) were taken for testing.

The training voice recordings of the same person was put into the same folder so that there will be 35 folders each containing 4 voice recording files. These folders were named as ‘voice1’, ‘voice2’, ‘voice3’ etc.
All the 70 test voice recording files were put into the same folder.

As there were no missing data or any data cleansing is required, the datasets are now ready to be trained and used for prediction.

CHAPTER 4 – SOLUTION DESIGN AND APPROACH

4.1 Overview

This chapter presents the architectural decisions taken for the solution and the rationale behind the decisions. A deep learning approach was used for the implementation, and it is described further in the chapter.

The below diagram shows the architecture of the deep learning-based implementation.

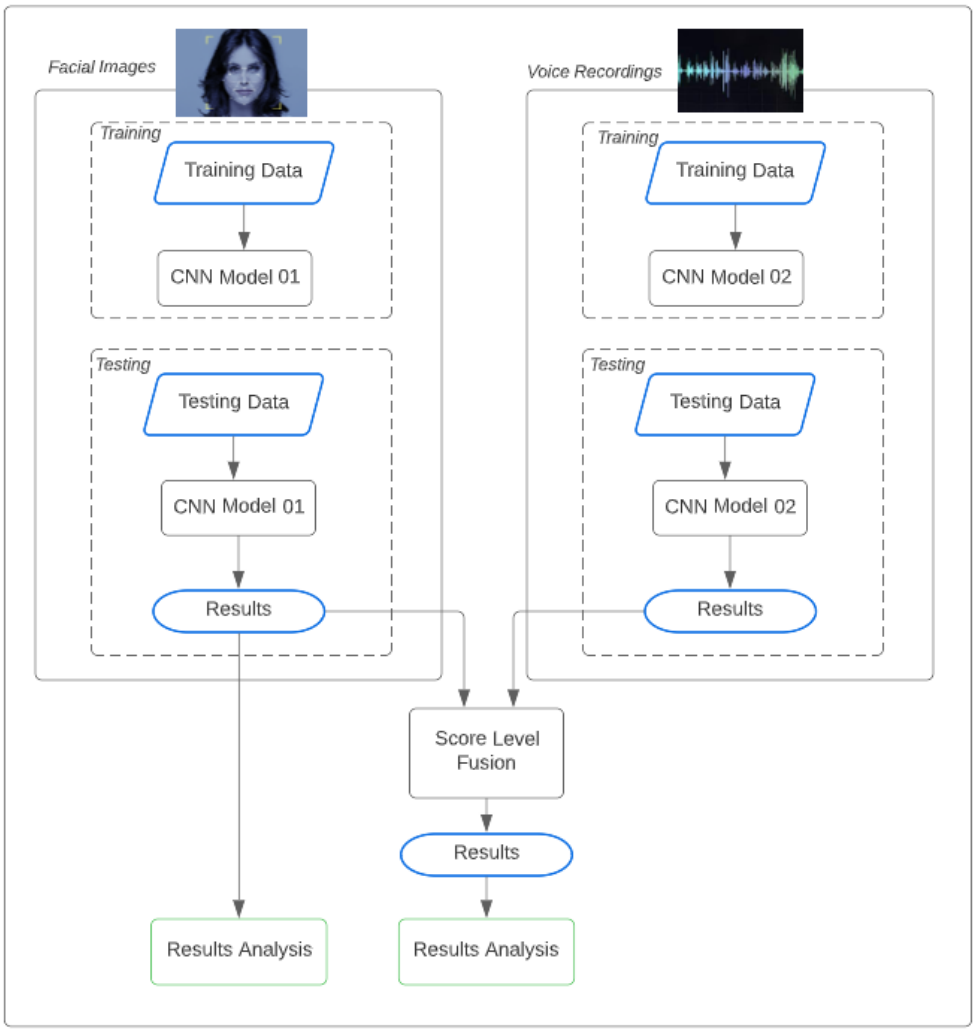

Figure 15: Proposed Deep Learning Approach Architecture

4.2 Deep Learning Approach based Architecture

4.2.1 Introduction to the Deep Learning Approach

Deep learning is a branch of machine learning which will add more 'depth' to the model transforming the data using various functions that allow data representation in a hierarchical way, through several levels of abstraction (Schmidhuber, 2015), (LeCun and Bengio, 1995).

A main advantage of deep learning is feature learning. It is capable of extracting features automatically from raw data, with features from higher levels of the hierarchy being formed by the composition of lower-level features (LeCun, Bengio and Hinton, 2015). Therefore, it does not require feature engineering whereas the other machine learning approaches rely on hand-engineered features, whose performance will affect heavily on the overall results. Moreover, it can solve complex problems particularly well and fast, because of more complex models used, which allow massive parallelization (Pan & Yang, 2010). These complex models employed in deep learning can increase classification accuracy or can reduce the error in regression problems. Even if deep learning is popular in numerous applications dealing with raster-based data (such as images and videos), it can also be applied to any form of data such as audio, speech and natural language (Kamilaris and Prenafeta-Boldú, 2018).

As described in **Error! Reference source not found.**, in the field of 2D face recognition, traditional m achine learning techniques have recently been superseded by deep learning techniques. Deep learning methods can be trained with large amounts of data to learn a face representation that is robust to the variations present in the training data set. Accordingly, they can learn from the training data instead of designing specialized features that are robust to different types of intra-class variations such as facial expression, age, illumination and pose. Other than learning these discriminative features, deep neural networks can also reduce the dimensionality and be trained as classifiers or using metric learning approaches (Kamilaris and Prenafeta-Boldú, 2018).

However, needing a very large dataset that has enough variations to generalize unseen samples is a major drawback of the deep learning methods. Also, it generally has a longer training time than the other machine learning methods (Kamilaris and Prenafeta-Boldú, 2018).

4.2.2 Convolutional Neural Networks (CNN)

As stated in the **Error! Reference source not found.**, Convolutional neural networks (CNNs) are t he most common type of deep learning method used for face recognition. It has recently seen major advancements and is one of the most common and effective applications in multiple image recognition tasks without limiting to face recognition.

CNN is a type of neural feed forward network consisting of multiple layers (Sharma et al., 2018; Paoletti et al., 2018; Liang et al., 2018). CNN consists of neurons or filters that have weights or parameters and bias that can be trained. The two main parts of a CNN are feature extraction and feature mapping (Liang et al., 2018).

The structure consists of convolutional layer, pooling, and fully connected layers. The objective of the convolution layer is to extract important features from input image data, which will result in an activation map or map of the output image (Zhu & Bain, 2017; Soltau et al., 2014). The pooling layer will reduce the dimensionality of each activation map generated by the convolution layer but will still have the most significant details. The fully connected layer is a fee-forward neural network. The entrance to a fully connected layer is the output of the final-pooling or convolution layer, which is flattened and then entered a fully connected layer (Nakahara et al., 2017).

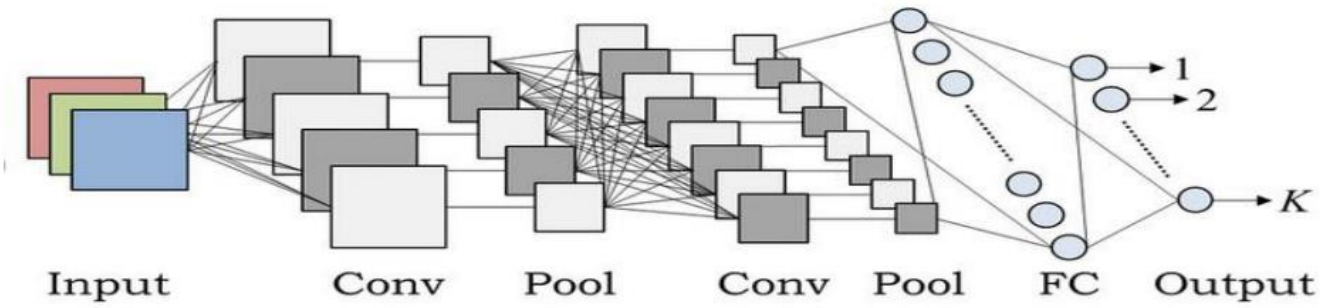

Figure 16: 1 Architecture of the CNN (Coskun et al., 2017)

Considering these factors and proven better performance on face recognition tasks, deep CNN was selected to use as the algorithm to first model.

Activation Function

An activation function in a neural network defines how the weighted sum of the input is transformed into an output from a node or nodes in a layer of the network. A few such activation functions are Rectified Linear Activation (ReLU), Logistic (Sigmoid) and Hyperbolic Tangent (Tanh). It is said that the ReLU function is the most common function used for hidden layers as it is simple to implement and effective at overcoming the limitations of other popular activation functions such as Sigmoid (Brownlee, 2021). It is also said that in modern neural networks, the default

recommendation is to use the rectified liner unit or ReLU activation function (Brownlee, 2021). Thus, decided to use the ‘ReLU’ activation function to train the CNN models.

Optimizers

In neural networks, optimizers are used to change the attributes of the neural network such as weights and learning rate to reduce the losses and thereby to provide the most accurate results possible. Stochastic Gradient Descent with Momentum (sgdm), Adaptive Moment Estimation (adam), Root Mean Square Propogation (rmsprop) and Adaptive Gradient Algorithm (Adagrad) are some examples of such optimizers. A study done to analyse the performance of different optimizers for deep learning-based image recognition has concluded that rmsprop has a better accuracy over sgdm and adam optimizers (Seda Postalcıoğlu, 2020). Another study done to classify Alzheimer disease (AD) has said that the accuracy of the CNN model using the rmsprop is 100% whereas the accuracy is 95.8% when adams optimizer is used (Taqi & Awad, 2018). Therefore, decided to use the rmsprop optimizer to train the CNN models.

4.2.3 Score Level Fusion

Fusion techniques are used to combine information from several biometric sources to improve the performance of the biometric recognition systems. These fusion techniques are of two major types as pre-classification and post-classification fusion techniques (Umer et al., 2020).

In pre-classification fusion techniques, the biometric information is fused before the classification task. Fusion techniques under this category are sensor-level and feature-level techniques. It is said that the performance of the pre-classification fusion techniques is prone to data redundancy, noise introduction and multi-environment image acquisition problems (Umer et al., 2020). Also, due to the great difference between face and voice features, it is difficult to fuse these feature vectors.

In post-classification fusion techniques, the biometric information is infused after the classification task based on the scores of the classifiers. It is said that the performance of post-classification fusion technique is free from noises and improves the recognition performance to a certain extent. Fusion techniques under this category are decision-level and matching score-level (Umer et al., 2020). The reliability of decision level fusion is lower than the other two as it refers only to the result of the unimodal biometric authentication, which will utilize less information about the original biometric characteristics. Matching score-level fusion considers different biometric features more sufficiently than the decision level (Umer et al., 2020).

The matching score-level fusion technique was employed in this study considering the above factors.

4.3 Model Construction

As described in the previous section, Convolutional Neural Networks (CNN) are capable of automatically detecting important features without human supervision. Therefore, after preparing the dataset, the CNN models were implemented.

Python 3.7.13 and the libraries Tensorflow 2.6.0 and Keras 2.6.0 were used to construct the face images-based CNN model and the voice recording based CNN model. Jupyter Notebook 6.4.8 was used as the IDE to contrast these models. The models were built as a sequence of layers by adding layers one at a time.

4.3.1 Face Image-based CNN Model

Main parameters of the image-based model building are as follows.

- Number of Nodes in the output layer = 35 and activate function = softmax
- The batch size = Number of samples processed before the model is updated o Batch size must be greater ≥ 1 or ≤ Number of samples in the data set
- Number of Epochs (Number of complete passes through the training dataset) = 15
- Activate function = Rectifier Activation Function (Relu)
- Loss = *categorical_crossentropy* as it is a multi-class classification
- Optimizer = Root Mean Square Propogation (rmsprop)
- Metrics = accuracy

The total time taken to train the image-based model was around 8 minutes.

```
In [12]: EndTime=time.time()
         print("###### Total Time Taken: ", round((EndTime-StartTime)/60), 'Minutes ######')

         ###### Total Time Taken:  8 Minutes ######
```

Figure 17: Total time taken to fit the image-based model

Preferably, we would like to have accuracy to be 1.0 (100%) and loss to be zero in the machine learning models. However, most machine learning solution does not always give 100 percent accuracy. Thus, the goal is to achieve highest accuracy and lowest loss in our data set. The below figure shows all the records of the model execution. Each of 15 epoch printing accuracy and loss of training data set as well as testing (validation) data set.

```
face_model = classifier.fit(training_set,epochs=15,validation_data=test_set,validation_steps=10)
```

```
Epoch 1/15
5/5 [==============================] - ETA: 0s - loss: 419.1810 - accuracy: 0.0500WARNING:tensorflow:Your input ran out of dat
a; interrupting training. Make sure that your dataset or generator can generate at least `steps_per_epoch * epochs` batches (in
this case, 10 batches). You may need to use the repeat() function when building your dataset.
5/5 [==============================] - 64s 15s/step - loss: 419.1810 - accuracy: 0.0500 - val_loss: 3.9715 - val_accuracy: 0.01
43
Epoch 2/15
5/5 [==============================] - 34s 7s/step - loss: 9.9701 - accuracy: 0.0429
Epoch 3/15
5/5 [==============================] - 35s 7s/step - loss: 8.4435 - accuracy: 0.0357
Epoch 4/15
5/5 [==============================] - 27s 6s/step - loss: 3.3906 - accuracy: 0.1214
Epoch 5/15
5/5 [==============================] - 27s 5s/step - loss: 3.2736 - accuracy: 0.1357
Epoch 6/15
5/5 [==============================] - 29s 6s/step - loss: 3.0246 - accuracy: 0.2286
Epoch 7/15
5/5 [==============================] - 29s 6s/step - loss: 3.0501 - accuracy: 0.2143
Epoch 8/15
5/5 [==============================] - 26s 5s/step - loss: 2.6297 - accuracy: 0.2429
Epoch 9/15
5/5 [==============================] - 27s 6s/step - loss: 2.6303 - accuracy: 0.3143
Epoch 10/15
5/5 [==============================] - 25s 5s/step - loss: 2.3674 - accuracy: 0.4500
Epoch 11/15
5/5 [==============================] - 30s 5s/step - loss: 2.4093 - accuracy: 0.3286
Epoch 12/15
5/5 [==============================] - 26s 5s/step - loss: 1.9711 - accuracy: 0.4643
Epoch 13/15
5/5 [==============================] - 29s 5s/step - loss: 1.9545 - accuracy: 0.4857
Epoch 14/15
5/5 [==============================] - 26s 5s/step - loss: 4.1208 - accuracy: 0.3143
Epoch 15/15
5/5 [==============================] - 28s 5s/step - loss: 4.5553 - accuracy: 0.3143
```

Figure 18: Epochs results of the image-based model

The summary of the trained image-based model is shown below. The figure shows the layers and their order, output shape and number of parameters (weights) in each layer and total number of parameters (weight) in the model.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 60, 60, 32)        2432
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 30, 30, 32)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 26, 26, 64)        51264
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 13, 13, 64)        0
_________________________________________________________________
flatten (Flatten)            (None, 10816)             0
_________________________________________________________________
dense (Dense)                (None, 64)                692288
_________________________________________________________________
dense_1 (Dense)              (None, 35)                2275
=================================================================
Total params: 748,259
Trainable params: 748,259
Non-trainable params: 0
_________________________________________________________________
```

Figure 19: Summary of the image-based model

4.3.2 Voice Recordings-based CNN Model

The training voice recordings were loaded to train the model and classified them as 35 classes. The *tf.audio.decode_wav* function of Tensorflow was used to read and decode the .wav audio files. The audio waves were then transformed into frequency domain using Tensorflow classes. Both a training and a validation dataset was prepared in this manner.

Thereafter, the voice recording-based model was trained as a separate feature engineering is not required. Similar to the previous model, epochs count was set to 15 to get a better accuracy.

- Number of Nodes in the output layer = 35 and activate function = softmax
- The batch size = Number of samples processed before the model is updated o Batch size must be greater ≥ 1 or ≤ Number of samples in the data set
- Number of Epochs (Number of complete passes through the training dataset) = 15
- Activate function = Rectifier Activation Function (Relu)
- Loss = sparse_*categorical_crossentropy*
- Optimizer = Root Mean Square Propogation (rmsprop)
- Metrics = accuracy

The time taken to train the voice recording-based model was less than 1 minutes.

```
EndTime=time.time()
print("###### Total Time Taken: ", round((EndTime-StartTime)/60), 'Minutes ######')

###### Total Time Taken:  0 Minutes ######
```

Figure 20: Time taken to train voice-based model

The below figure shows all the records of the model execution along with the accuracy and loss values in each epoch.

```
voice_model = model.fit(train_ds, epochs=15, validation_data=valid_ds, validation_steps=10)

Epoch 1/15
2/2 [==============================] - ETA: 0s - loss: 15.3311 - accuracy: 0.0214WARNING:tensorflow:Your input ran
interrupting training. Make sure that your dataset or generator can generate at least `steps_per_epoch * epochs` ba
is case, 10 batches). You may need to use the repeat() function when building your dataset.
2/2 [==============================] - 3s 1s/step - loss: 15.3311 - accuracy: 0.0214 - val_loss: 57.2509 - val_accu
Epoch 2/15
2/2 [==============================] - 1s 233ms/step - loss: 57.6495 - accuracy: 0.1214
Epoch 3/15
2/2 [==============================] - 1s 208ms/step - loss: 31.3292 - accuracy: 0.1929
Epoch 4/15
2/2 [==============================] - 2s 237ms/step - loss: 27.8159 - accuracy: 0.2071
Epoch 5/15
2/2 [==============================] - 2s 215ms/step - loss: 11.1333 - accuracy: 0.3071
Epoch 6/15
2/2 [==============================] - 1s 235ms/step - loss: 21.1725 - accuracy: 0.2214
Epoch 7/15
2/2 [==============================] - 2s 247ms/step - loss: 10.3442 - accuracy: 0.3714
Epoch 8/15
2/2 [==============================] - 1s 223ms/step - loss: 5.1175 - accuracy: 0.4929
Epoch 9/15
2/2 [==============================] - 1s 250ms/step - loss: 5.9350 - accuracy: 0.4571
Epoch 10/15
2/2 [==============================] - 2s 217ms/step - loss: 4.4145 - accuracy: 0.5500
Epoch 11/15
2/2 [==============================] - 2s 230ms/step - loss: 8.9194 - accuracy: 0.4214
Epoch 12/15
2/2 [==============================] - 2s 259ms/step - loss: 3.6664 - accuracy: 0.5571
Epoch 13/15
2/2 [==============================] - 2s 248ms/step - loss: 2.0622 - accuracy: 0.6357
Epoch 14/15
2/2 [==============================] - 2s 274ms/step - loss: 1.3478 - accuracy: 0.7214
Epoch 15/15
2/2 [==============================] - 2s 246ms/step - loss: 0.9886 - accuracy: 0.7571
```

Figure 21: Epochs results of the voice-based model

The summary of the fitted voice-based model is as follows:

```
model.summary()
```

```
Model: "model"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input (InputLayer)              [(None, 8000, 1)]    0
__________________________________________________________________________________________________
conv1d_1 (Conv1D)               (None, 8000, 16)     64          input[0][0]
__________________________________________________________________________________________________
conv1d (Conv1D)                 (None, 8000, 16)     32          input[0][0]
__________________________________________________________________________________________________
add (Add)                       (None, 8000, 16)     0           conv1d_1[0][0]
                                                                 conv1d[0][0]
__________________________________________________________________________________________________
max_pooling1d (MaxPooling1D)    (None, 4000, 16)     0           add[0][0]
__________________________________________________________________________________________________
flatten_1 (Flatten)             (None, 64000)        0           max_pooling1d[0][0]
__________________________________________________________________________________________________
dense_2 (Dense)                 (None, 256)          16384256    flatten_1[0][0]
__________________________________________________________________________________________________
output (Dense)                  (None, 35)           8995        dense_2[0][0]
==================================================================================================
Total params: 16,393,347
Trainable params: 16,393,347
Non-trainable params: 0
__________________________________________________________________________________________________
```

Figure 22: Summary of the voice-based model

4.4 Prediction

The trained face image-based model and the voice recording-based models were tested on the unseen and face images and voice recordings, respectively.

Both the probability values predicted for each face/ voice and the final predicted face/ voice class (using *np.argmax* function) were saved for further analysis of the prediction.

CHAPTER 5 – EVALUATION & RESULTS ANALYSIS

5.1 Overview

In this chapter, the evaluation that will be carried out to assess the performance and the validity of the model designed to improve the performance of face recognition using voice acoustics during the COVID-19 era are presented. The metrics that will be used to evaluate the proposed models are discussed in detail at the beginning of the Chapter.

5.1 Evaluation Plan

There are different approaches that can be utilized to evaluate a research study such as mathematical proof, experimental-based, Opinion & Interview based etc. Out of these different approaches, an experiment-based approach is applied in this study to measure and evaluate the performance of the models.

As described in 2.5 Evaluation Metrics, most of the studies have used a publicly available dataset to evaluate the performance of face recognition systems and bimodal person identification systems so that the performance can be compared with the other systems proposed for the same purpose. However, as there are no public datasets with masked face images and masked voice recording, a public dataset cannot be used in this study and therefore the performance cannot be compared with the other similar systems/ models. Thus, a private dataset which was collected as mentioned in

3.3 Data Collection is used to evaluate the performance.

Moreover, *Anaconda Jupyter Notebook* is used to calculate the values of the selected evaluation metrics.

5.2 Evaluation Metrics

The performance of the models is measured by treating the classification as a binary classification (i.e., by considering one class vs. all other classes). As described in 2.5 Evaluation Metrics, in literature review it was found that most of the similar studies have used accuracy as the evaluation metric. Thus, accuracy is the main evaluation metric in this study as well. In addition, recall and precision values are also measured. For the convenience, we have used Ci to refer to a person.

Accuracy, Precision & Recall

A confusion matrix or a contingency table consists of information on actual vs. predicted classifications in a machine learning system. Confusion matrices can be drawn separately for each class Ci as shown in the table.

		Predicted	
		Class C_i	Not Class C_i
Actual	Class C_i	**TP**	**FN**
	Not class C_i	**FP**	**TN**

Table 2: Confusion matrix

The correctness of these classifications can be evaluated by computing the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values for each class.

- True Positive (TP): Declare the class as C_i when the given class is C_i
- False Negative (FN): Not declaring the class as C_i, when the given class is C_i
- False Positive (FP): Declare the class as C_i when it is not given as C_i
- True Negative (TN): Not declaring the class as C_i, when it is not given as C_i

Accuracy measure shows the overall effectiveness of the predictions for each class Ci. Accuracy measure of the class Ci is calculated as:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

Precision (Specificity) measure gives the proportion of number of variants correctly identified as those belong to class Ci, to the total number of variants identified as class Ci.

$$Precision = \frac{TP}{TP + FP}$$

Recall (Sensitivity) measure gives the proportion of correctly identifying variants as those belong to class C_i.

$$Recall = \frac{TP}{TP + FN}$$

The overall performance of the model can be evaluated by calculating the accuracy. Recall and precision values can be calculated per class, and then can be averaged over all available classes as it has been done in previous studies. I.e., the macro-average recall and precision values can be calculated.

Accordingly, the accuracy, precision and recall values calculated for the unimodal where only face images are used and for the bimodal where both the face images and voice recordings are used. In this way, the performance of each model can be compared.

5.3 Evaluation of the Uni-Models

Accuracy, macro-averaged precision, and macro-averaged recall values obtained for the image-based model alone are 0.1429, 0.1184 and 0.1429 respectively.

```
from sklearn.metrics import accuracy_score, precision_score, recall_score
accuracy_image_model = accuracy_score(actual_values_images, predicted_values_images)
precision_image_model = precision_score(actual_values_images, predicted_values_images, average='macro')
recall_image_model = recall_score(actual_values_images, predicted_values_images, average='macro')


print('Accuracy of the image-based unimodal is ' , accuracy_image_model)
print('Precision of the image-based unimodal is ' , precision_image_model)
print('Recall of the image-based unimodal is ' , recall_image_model)
```

```
Accuracy of the image-based unimodal is  0.14285714285714285
Precision of the image-based unimodal is  0.11845722904546435
Recall of the image-based unimodal is  0.14285714285714285
```

Figure 23: Accuracy, Precision & Recall of Image-based Model

Accuracy, macro-averaged precision, and macro-averaged recall values obtained for the voice-based model alone are 0.0143, 0.0015 and 0.0143 respectively.

```
addition = face_pred_probabilites + y_pred
average = addition / 2
average
```

```
array([[1.40295888e-02, 2.75746326e-02, 1.04907927e-03, ...,
        1.72850618e-03, 1.84145545e-05, 2.81321682e-01],
       [4.31989643e-03, 1.68988362e-01, 1.19212451e-02, ...,
        2.32192341e-02, 2.58273806e-03, 9.77287306e-03],
       [1.74322760e-02, 4.50591775e-04, 1.72856399e-02, ...,
        9.17134900e-03, 2.37567601e-04, 2.14391169e-02],
       ...,
       [5.38386440e-03, 2.36817026e-03, 2.24284344e-02, ...,
        1.46267637e-01, 6.71416122e-02, 3.74942492e-03],
       [3.95781127e-03, 7.24446834e-02, 4.54444486e-03, ...,
        5.43966220e-02, 1.87036465e-03, 8.76446720e-05],
       [1.08987085e-02, 1.17674598e-02, 9.55965884e-03, ...,
        8.01532700e-02, 3.11694330e-02, 2.03553810e-02]])
```

```
predicted_persons = []
for i in average:
    predicted_person = np.argmax(i)
    predicted_persons.append(predicted_person)
```

```
accuracy_voice_model = accuracy_score(test_labels, predicted_speaker)
precision_voice_model = precision_score(test_labels, predicted_speaker, average='macro')
recall_voice_model = recall_score(test_labels, predicted_speaker, average='macro')


print('Accuracy of the voice-based unimodal is ' , accuracy_voice_model)
print('Precision of the voice-based unimodal is ' , precision_voice_model)
print('Recall of the voice-based unimodal is ' , recall_voice_model)
```

```
Accuracy of the voice-based unimodal is  0.014285714285714285
Precision of the voice-based unimodal is  0.0015037593984962405
Recall of the voice-based unimodal is  0.014285714285714285
```

Figure 24: Accuracy, Precision & Recall of Voice-based Model

5.5 Evaluation of the Bi-Models

A score level-based fusion described in 4.2.3 Score Level Fusion was performed to combine the results of the uni-models. This was done by taking the average of probability values predicted by each model for each of the test images and voices. Then the predicted person was obtained by taking the person corresponding to the maximum averaged probability values.

Figure 25: Score level fusion of the uni-model results

Thereafter, the accuracy, recall and precision values of the bimodels were calculated using the score level-based results. The accuracy, macro-averaged precision and macro-averaged recall values obtained for the combined models were 0.0286, 0.0157 and 0.0278.

```
accuracy_combined_models = accuracy_score(test_labels, predicted_persons)
precision_combined_model = precision_score(test_labels, predicted_persons, average='macro')
recall_combined_model = recall_score(test_labels, predicted_persons, average='macro')

print('Accuracy of the combined bi-modal is ' , accuracy_combined_models)
print('Precision of the combined bi-imodal is ' , precision_combined_model)
print('Recall of the combined bi-modal is ' , recall_combined_model)
```

```
Accuracy of the combined bi-modal is  0.02857142857142857
Precision of the combined bi-imodal is  0.01574074074074074
Recall of the combined bi-modal is  0.027777777777777776
```

Figure 26: Accuracy, Precision & Recall of Bimodals

CHAPTER 6 – CONCLUSION & FUTURE WORK

6.1 Overview

This chapter describes the conclusion of the research study and the future work that can be done to improve it.

6.2 Conclusion

A comparison of the accuracy, precision and recall values obtained for each model individual and for the combined models are as follows.

Table 3: Accuracy, Precision & Recall of All Models

	Accuracy (%)	**Precision (%)**	**Recall (%)**
Image-based Model	14.29	11.84	14.29
Voice-based Model	01.43	00.15	01.43
Combined Model	02.86	01.57	02.78

Although the accuracy, precision and recall values of the image-based model is comparatively high, the values obtained for the combined models are low as the results of the voice-based model are low.

There can be several reasons for the lower performance of the model. One main reason could be the amount of data used to train the CNN models. Usually, as described in 4.2.1 Introduction to the Deep Learning Approach, a major drawback of the deep learning approaches is that it requires a large amount of data. For this study, images and voice recordings were collected only from a small sample of size 35. Therefore, the data used for training might not have been enough for the models to perform better.

A reason for the voice-baaed model to have a lower performance could be the lack of pre-processing the voice recordings by removing noise or by training the model to learn and avoid noise in the captured recordings. Moreover, although it has been proved in some past studies that the some of the voice characteristics will be the same regardless of if a face mask is wear or not, it could be that most of the voice characteristics might change depending on if a face mask is there or not.

Thus, it can be concluded that incorporating voice features to the face features for person identification during a pandemic with a small sample data will not improve the performance of person identification.

6.3 Future Work

There are several other approaches such as Support Vector Mechanism (SVM) that does not require a large amount of data as it is required for deep learning approaches. Thus, such an approach can be tried out on the same dataset to evaluate which approach has a better performance. Moreover, more data can be collected to evaluate the performance of the CNN models.

Further, as in this study, only the facial features and acoustic voice features were considered when wearing a face mask, the other biometric features such as fingerprint or iris or palm print or signature etc, can also be extracted and combined with the facial features to evaluate which biometric feature combination has a better performance.

REFERENCES

Lawton, G., Le Page, M., Vaughan, A. and Wilson, C., 2021. How do we live with covid-19?. *New Scientist*, 251(3344), pp.8-11.

Wu, Y., Chen, C. and Chan, Y., 2020. The outbreak of COVID-19: An overview. *Journal of the Chinese Medical Association*, 83(3), pp.217-220.

Eikenberry, S., Mancuso, M., Iboi, E., Phan, T., Eikenberry, K., Kuang, Y., Kostelich, E. and Gumel, A., 2020. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. Infectious Disease Modelling, 5, pp.293-308.

Abboah-Offei, M., Salifu, Y., Adewale, B., Bayuo, J., Ofosu-Poku, R. and Opare-Lokko, E., 2021. A rapid review of the use of face mask in preventing the spread of COVID-19. International Journal of Nursing Studies Advances, 3, p.100013.

Wang, C., Ng, C. and Brook, R., 2020. Response to COVID-19 in Taiwan. JAMA, 323(14), p.1341.
Bricker, D., 2022. [online] Available at: <https://www.ipsos.com/en/more-people-say-theyre-wearing-masks-protect-themselves-covid-19-march> [Accessed 3 June 2021].

Abboah-Offei, M., Salifu, Y., Adewale, B., Bayuo, J., Ofosu-Poku, R. and Opare-Lokko, E., 2021. A rapid review of the use of face mask in preventing the spread of COVID-19. International Journal of Nursing Studies Advances, 3, p.100013.

Coronavirus Disease 2019 (COVID-19). 2022. Centers for Disease Control and Prevention. [online] Available at: <https://www.cdc.gov/coronavirus/2019-ncov/index.html> [Accessed 3 June 2022].

Abboah-Offei, M., Salifu, Y., Adewale, B., Bayuo, J., Ofosu-Poku, R. and Opare-Lokko, E., 2021. A rapid review of the use of face mask in preventing the spread of COVID-19. International Journal of Nursing Studies Advances, 3, p.100013.

Kortli, Y., Jridi, M., Al Falou, A. and Atri, M., 2020. Face Recognition Systems: A Survey. Sensors, 20(2), p.342.

Parmar, D. and Mehta, B., 2022. Face Recognition Methods & Applications. [online] arXiv.org. Available at: <https://arxiv.org/abs/1403.0485> [Accessed 3 June 2021].

Hariri, W., 2021. Efficient masked face recognition method during the COVID-19 pandemic. Signal, Image and Video Processing, 16(3), pp.605-612.

Ngan, M., Grother, P. and Hanaoka, K., 2022. Ongoing Face Recognition Vendor Test (FRVT) Part 6B: Face recognition accuracy with face masks using post-COVID-19 algorithms. [online] NIST. Available at: <https://www.nist.gov/publications/ongoing-face-recognition-vendor-test-frvt-part-6b-face-recognition-accuracy-face-masks> [Accessed 3 June 2021].

Damer, N., Boutros, F., Süßmilch, M., Fang, M., Kirchbuchner, F. and Kuijper, A., 2022. Masked face recognition: Human versus machine. IET Biometrics,.

Wilmer, J., 2017. Individual Differences in Face Recognition: A Decade of Discovery. Current Directions in Psychological Science, 26(3), pp.225-230.

Zhao, W., Chellappa, R., Phillips, P. and Rosenfeld, A., 2003. Face recognition. ACM Computing Surveys, 35(4), pp.399-458.

Adjabi, I., Ouahabi, A., Benzaoui, A. and Taleb-Ahmed, A., 2020. Past, Present, and Future of Face Recognition: A Review. Electronics, 9(8), p.1188.

Boyko, N., Basystiuk, O. and Shakhovska, N., 2022. Performance Evaluation and Comparison of Software for Face Recognition, Based on Dlib and Opencv Library. [online] Available at: <https://www.researchgate.net/publication/328087597_Performance_Evaluation_and_Comparison_of_Software_for_Face_Recognition_Based_on_Dlib_and_Opencv_Library> [Accessed 3 June 2021].

Azeem, A., Raza, M., Sharif, M. and Murtaza, M., 2013. A Survey: Face Recognition Techniques under Partial Occlusion. The International Arab Journal of Information Technology, [online] 11(1). Available at: <https://iajit.org/portal/PDF/vol.11,no.1/4135.pdf> [Accessed 3 June 2021].

Damer, N., Henry Grebe, J. and Chen, C., 2020. The Effect of Wearing a Mask on Face Recognition Performance: an Exploratory Study. IEEE, [online] Available at: <https://ieeexplore.ieee.org/document/9210999> [Accessed 3 June 2021].

Wang, Z., Wang, G., Huang, B., Xiong, Z., Hong, Q., Wu, H., Yi, P., Jiang, K., Wang, N., Pei, Y., Chen, H., Miao, Y., Huang, Z. and Liang, J., 2020. Masked Face Recognition Dataset and Application. [online] arXiv.org. Available at: <https://arxiv.org/abs/2003.09093> [Accessed 3 June 2021].

Vu, H., Nguyen, M. and Pham, C., 2021. Masked face recognition with convolutional neural networks and local binary patterns. Applied Intelligence, 52(5), pp.5497-5512.

Alzu'bi, A., Albalas, F., AL-Hadhrami, T., Younis, L. and Bashayreh, A., 2021. Masked Face Recognition Using Deep Learning: A Review. Electronics, 10(21), p.2666.

Phys.unsw.edu.au. 2022. Voice Acoustics: an introduction to the science of speech and singing. [online] Available at: <http://www.phys.unsw.edu.au/jw/voice.html> [Accessed 3 June 2022].

Nguyen, D., McCabe, P., Thomas, D., Purcell, A., Doble, M., Novakovic, D., Chacon, A. and Madill, C., 2021. Acoustic voice characteristics with and without wearing a facemask. Scientific Reports, 11(1).

En.wikipedia.org. 2022. Convenience sampling - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/Convenience_sampling> [Accessed 3 June 2022].

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Networks, 61, pp.85-117.

LeCun, Y. and Bengio, Y., 1995. [online] Yann.lecun.com. Available at: <http://yann.lecun.com/exdb/publis/pdf/lecun-bengio-95a.pdf> [Accessed 3 June 2022].

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. Nature, 521(7553), pp.436-444.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on knowledge and data engineering , 22(10), 1345-1359.

Kamilaris, A. and Prenafeta-Boldú, F., 2018. Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, pp.70-90.

Sharma, N., Jain, V., & Mishra, A. (2018). An Analysis Of Convolutional Neural Networks For Image Classification. Procedia Computer Science, 132, 377–384. https://doi.org/10.1016/j.procs.2018.05.198

Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. (2018). A new deep convolutional neural network for fast hyperspectral image classification. ISPRS Journal of Photogrammetry and Remote Sensing, 145, 120–147. https://doi.org/10.1016/j.isprsjprs.2017.11.021.

Liang, G., Hong, H., Xie, W., & Zheng, L. (2018). Combining Convolutional Neural Network With Recursive Neural Network for Blood Cell Image Classification. IEEE Access, 6, 36188–36197. https://doi.org/10.1109/ACCESS.2018.2846685.

Coskun, M., Ucar, A., Yildirim, O., & Demir, Y. (2017). Face recognition based on convolutional neural network. 2017 International Conference on Modern Electrical and Energy Systems (MEES), 376–379. https://doi.org/10.1109/MEES.2017.8248937.

Zhu, X., & Bain, M. (2017). B-CNN: Branch Convolutional Neural Network for Hierarchical Classification. ArXiv:1709.09890 [Cs]. http://arxiv.org/abs/1709.09890.

Soltau, H., Saon, G., & Sainath, T. N. (2014). Joint training of convolutional and non-convolutional neural networks. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5572–5576. https://doi.org/10.1109/ICASSP.2014.6854669.

Nakahara, H., Fujii, T., & Sato, S. (2017). A fully connected layer elimination for a binarizec convolutional neural network on an FPGA. 2017 27th International Conference on Field Programmable Logic and Applications (FPL), 1–4. https://doi.org/10.23919/FPL.2017.8056771.

Umer, S., Sardar, A., Dhara, B., Rout, R. and Pandey, H., 2020. Person identification using fusion of iris and periocular deep features. Neural Networks, 122, pp.407-419.

P. S, S. and J. B, P., 2013. An Overview of Multimodal Biometrics. Signal & Image Processing: An International Journal, 4(1), pp.57-64.

Kartik, P., Vara Prasad, R. and Mahadeva Prasanna, S., 2008. Noise robust multimodal biometric person authentication system using face, speech and signature features. IEEE.

S. Hariprasath and T. Prabakar, “Multimodal Biometric Recognition using Iris Feature Extraction and Palmprint Features,” in Proc. of International Conference on Advances in Engineering, Science and Management (ICAESM), Nagapattinam, pp. 174-179, 30-31 March 2012.

T. Murakami and K. Takahashi, “Fast and Accurate Biometric Identification Using Score Level Indexing and Fusion,” in Proc. Of International Joint Conference on Biometrics (IJCB), USA, ,pp. 978-985, 2011.

Carolinewright.com. 2022. The Rainbow Passage | Caroline Wright – Artist, England. [online] Available at: <http://www.carolinewright.com/portfolio/the-rainbow-passage/#:~:text=The%20Rainbow%20Passage%20is%20a,test%20breathing%20and%20speech%20patterns.> [Accessed 3 June 2022].

Asha.org. 2022. Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ASHA Special Interest Group 3, Voice and Voice Disorders. [online] Available at: <https://www.asha.org/siteassets/uploadedFiles/ASHA/SIG/03/CAPE-V-Procedures-and-Form.pdf> [Accessed 3 June 2022].

McLaughlin, N., Ming, J. and Crookes, D., 2022. Robust Multimodal Person Identification with Limited Training Data. IEEE Transactions on Human Machine Systems, 43(2), pp.214 - 224.

Soleymani, S., Dabouei, A., Kazemi, H. and Dawson, J., 2022. Multi-Level Feature Abstraction from Convolutional Neural Networks for Multimodal Biometric Identification. 2018 24th International Conference on Pattern Recognition (ICPR),.

Cherrat, E., Alaoui, R. and Bouzahir, H., 2020. Convolutional neural networks approach for multimodal biometric identification system using the fusion of fingerprint, finger-vein and face images. PeerJ Computer Science,.

Postalcıoğlu, S. (no date) Performance analysis of different optimizers for deep learning-based image recognition, International Journal of Pattern Recognition and Artificial Intelligence. Available at: https://www.worldscientific.com/doi/10.1142/S0218001420510039 (Accessed: November 21, 2022).

Taqi, A.M. and Awad, A. (2018) The impact of multi-optimizers and data augmentation on tensorflow convolutional neural network performance, IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/8396988/ (Accessed: November 21, 2022).

Brownlee, J. (2021) How to choose an activation function for deep learning, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/ (Accessed: November 21, 2022).

APPENDIX A - INFORMED CONSENT LETTER

TITLE OF STUDY: Acoustic Voice Characteristics for Face Recognition during the Covid-19 Pandemic

PRINCIPAL INVESTIGATOR:

Dakshila Kamalsooriya,

University of Colombo School of Computing,

35 Reid Avenue, Colombo 07.
071-2596506
ndakshila@gmail.com

PURPOSE OF STUDY:

You are being asked to take part in a research study. Before you decide to participate in this study, it is important that you understand why the research is being done and what it will involve. Please read the following information carefully. Please ask the researcher if there is anything that is not clear or if you need more information. You may terminate your involvement at any time if you choose.

The purpose of this study is to find out if voice acoustic characteristics can be used to improve the performance of face recognition when wearing a face mask to enable contactless and smooth-running operations and thereby help to curb the transmission of the COVID-19 virus from one person to another.

STUDY PROCEDURES:

- Facial images and voice recordings should be collected for the study both with and without facial masks.
- After pre-processing the collected data, features will be extracted from the collected images and voice recordings.
- These features will then be used to train models. Only the facial features will be used to train one model. To train the other models, both the facial features and voice features will be used.
- The trained models will then be used to predict the testing data set.
- Predicted results will be used to analyze the performance of the models.

To capture facial images and voice recordings both with and without masks, the participants should follow the instructions below.

Capturing facial images

- Provided Canon EOS 300d camera should be used with its default settings.
- The images should be taken in a white background. The distance to the camera should be around 6 feet and it should be placed at the eye level.
- N95 mask should be worn to take images with masks.
- 3 images each with and without wearing mask should be taken.

Capturing voice recordings

- Provided Sony Digital Voice Recorder ICD PX 470 should be used with its default settings.
- The images should be taken in a white background. The distance to the voice recorder should be around 6 feet and it should be placed at the eye level.
- N95 mask should be worn to take voice recordings with masks.
- The following tasks should be read to take voice recordings.
    - Sustain voicing of the vowel /a/, at a comfortable pitch and loudness levels for at least 10 seconds after a deep breath.
    - The Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) – 3rd phrase (CAPEV-3).
    - The Rainbow Passage - 2nd and 3rd sentences.

CONFIDENTIALITY:

Participant data will be kept confidential except in cases where the researcher is legally obligated to report specific incidents. The data will be destroyed once the study is performed.

VOLUNTARY PARTICIPATION:

Your participation in this study is voluntary. It is up to you to decide whether to take part in this study or not. If you decide to take part in this study, you will be asked to sign this consent form. After you sign the consent form, you are still free to withdraw at any time and without giving a reason. Withdrawing from this study will not affect the relationship you have, if any, with the researcher. If you withdraw from the study before data collection is completed, your data will be returned to you or destroyed.

CONSENT:

I have read, and I understand the provided information and have had the opportunity to ask questions. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason and without cost. I understand that I will be given a copy of this consent form. I voluntarily agree to take part in this study.

Participant's signature _____ Date _____

Investigator's signature _____ Date _____