Image analysis for a mobile phone-based assessment of latent tuberculosis infection.

UNIVERSITY OF CAPE TOWN
Department of Human Biology
Division of Biomedical Engineering

Dissertation
MSc Biomedical Engineering

Sarah Maclean | MCLSAR002

September 2020

Supervisor: Dr Tinashe Mutsvangwa

Co-supervisors: Dr Bessie Malila, Prof Tania Douglas

University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

University of Cape Town

Declaration

I, Sarah Maclean, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: Signed by candidate

Date: 2020/09/11

Abstract

The current, most widely used method to screen for latent tuberculosis infection is the Mantoux tuberculin skin test, where tuberculin is injected into a patient's arm and may result in a cutaneous induration forming at the site of injection. A diameter measurement of the resultant induration, recorded using a ruler and ball point pen, is currently used to indicate the presence of latent tuberculosis infection. Limitations associated with the tuberculin skin test procedure are the crudeness of the induration measurement method, the follow-up clinical visit required from patients to have their induration measured, and the need for trained clinicians who can perform the induration measurement. These limitations motivated research into a mobile phone-based screening system which can be used to obtain a more accurate measurement of the induration without the need for a second visit to the clinic by patients. The prototype screening tool consists of a user interface for capturing induration images and a backend processing system that produces a three-dimensional reconstruction of the induration for measurement. Recommendations from previous studies on the prototype screening tool, which involved evaluation of the mobile application using mock induration images, included improving the accuracy of measuring the induration and evaluating the tool on real induration images. The aim of this study was to evaluate the existing backend system and explore an alternative assessment approach for assessing the induration. This was achieved through the following objectives: (1) applying the current backend system to real induration images, (2) examining the need for three-dimensional reconstruction for delineation of the induration for measurement and (3) exploring an alternative method for the assessment of induration images using deep learning. Results for the first objective showed the three-dimensional reconstruction to be unsuccessful on real images. This was due to the homogeneity between the indurations and the surrounding skin, rendering the algorithm ineffective in delineating the indurations to obtain the diameter measurement required for diagnosis. The second objective involved determining whether the image orientation or induration height affected the diagnostic measurement. It was found that real indurations are much flatter and more subtle compared to the mock indurations used in the previous studies. This motivated an alternative image assessment approach using deep learning. However, deep learning approaches require large databases of annotated images to prevent overfitting on training data. The last objective therefore involved the design and implementation of a generative adversarial network for generation of synthetic images from a limited number of real images, which allowed the generation of an unlimited number of realistic-looking synthetic images from 150 real induration images.

Acknowledgements

I would like to express my sincere thanks and gratitude to the following persons:

My supervisor, Dr Tinashe Mutsvangwa, for his constant support, encouragement and phenomenal knowledge in the field of Biomedical Engineering. His insightful advice together with his persistence to encourage hard work has benefited me beyond my research at the University of Cape Town.

My co-supervisors, Prof Tania Douglas and Dr Bessie Malila, for their selfless support throughout the course of producing this dissertation and for their very valuable advice and suggestions.

My colleagues in the Medical Image Inferencing and Distributed Diagnostics (Mi2D2) group and the Health Innovation office for their great friendship. The laughter and kindness they brought to my work environment made my research experience very enjoyable.

The staff and my friends from the Division of Biomedical Engineering for creating such a friendly and stimulating work atmosphere.

My family, Llewellyn, Christine, Morgan and Ty Maclean as well as Ryan Thompson for their love and unwavering support throughout my studies.

Thanks again to Prof Tania Douglas and Dr Tinashe Mutsvangwa for choosing me to take up this project and for the sponsorship of my research through the DST/NRF South African Research Chairs Initiative which provided support to me through Prof Tania Douglas's Research Chair in Biomedical Engineering and Innovation (grant number 98788).

Contents

List of Figures and Tables

Abbreviations

APS	Agisoft PhotoScan
CNN	Convolutional Neural Network
DCDAN	Deep Convolutional Generative Adversarial Network
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
HIV	Human Immunodeficiency Virus
HREC	Human Research Ethics Committee
IGRA	Interferon-Gamma Release Assay
IS	Inception Score
LTBI	Latent Tuberculosis Infection
RGB	Red Green Blue
SFM	Structure-From-Motion
SWD	Sliced Wasserstein Distance
TB	Tuberculosis
TST	Tuberculin Skin Test
UCT	University of Cape Town
WHO	World Health Organisation

1 Introduction

One of the United Nations Sustainable Development Goals is to ensure health for all at all ages by 2030 (UNSDG, 2018). Tuberculosis (TB) has been reported as the leading cause of death from an infectious disease worldwide, consequently threatening to derail the efforts of national governments in achieving this goal (WHO, 2017). Individuals who have latent tuberculosis infection (LTBI) have Mycobacterium tuberculosis (M.tuberculosis) bacteria in their blood stream, but do not present with any of the common symptoms of active TB. Although LTBI is an asymptomatic condition, it enables active tuberculosis to develop in individuals with compromised immune systems. The most common comorbidity to TB is the human immunodeficiency virus (HIV) (Manosuthi, Wiboonchutikul & Sungkanuparph, 2016). The development of active TB from LTBI can be prevented by TB preventative medication, but administration of the medication is reliant on an accurate LTBI diagnosis (WHO, 2017).

The current, most widespread method used to screen for LTBI is the Mantoux tuberculin skin test (TST) (WHO, 2017). During the TST, tuberculin is injected into the patient's forearm. A cutaneous induration forms at the site of injection if there are M. tuberculosis bacteria present in the patient's blood stream. The diameter of the resultant induration is used to indicate the presence of LTBI and is generally measured by health professionals using the ballpoint pen and ruler method (WHO, 2017). The method involves marking the edges of the induration using a ballpoint pen and the diameter is then measured using a ruler (Al-Orainey, 2009). The crudeness of the ballpoint pen and ruler method makes the measurement prone to error (Meyer, Hougen & Edwards, 1951). The placement of the ruler during measurement often results in deformation of the induration leading to inaccurate readings. Substandard precision also arises in the pen markings, due to their exact positioning being left up to the discretion of the clinician.

An additional drawback linked to the traditional TST measurement method is the need for follow-up visits 48 to 72 hours after administration of the tuberculin (Haghdoost et al., 2014). The follow-up visit is required for measurement of the induration when it is at peak diameter (Morán-Mendoza et al., 2007). Patients often do not return to the clinic to have their induration measurement taken due to several reasons such as forgetting, other commitments or the cost involved with travelling (Risser et al., 1985). Furthermore, there are a limited number of skilled clinicians who are able to perform

the induration measurement. As a result, in some cases inexperienced people do the measurements, which increases the possibility of misdiagnosis (Siroka et al., 2016).

The limitations associated with this traditional TST measurement procedure motivated research into the use of mobile technologies for LTBI screening in the Division of Biomedical Engineering at the University of Cape Town (UCT) (Dendere et al., 2017). The current prototype of the UCT mobile phone-based system consists of a user interface for image capture, and a backend image analysis system for evaluation of the images and result generation (Naraghi et al., 2018). These subsystems work together to perform an evaluation of the TST induration response. However, there is still a need to improve the usability and diagnostic accuracy of the mobile health (mHealth) system.

1.1 Problem Statement and Motivation

While the performance of the current mobile phone-based LTBI screening tool is promising, drawbacks associated with the current system have motivated further research into improvement of the backend processing system.

The first drawback is that a high number of images (7) of the induration are required to be sent from a mobile phone to a remote processing centre. The high number stems from the number of images required to accurately reconstruct the induration in three dimensions (3D) before calculating the diameter measurement. Reduction of the number of images used during the assessment is thus necessary to reduce the inconvenience associated with image capture, increase the usability of the application, and decrease the data costs associated with image transmission.

Secondly, the algorithms used to delineate the induration require improvement to increase the accuracy of the diameter measurement.

Lastly, the mHealth system has only been evaluated on simulated indurations. It is therefore necessary for the application to be evaluated in a clinical setting using real indurations. However, assessment of the application requires a large number of real images in order for the results to be statistically robust. To date, however, only a limited number of real induration images are available for use in testing the application. This is a common limitation in medical image analysis, regardless of modality. Limitations related to cost, privacy and ethical barriers need to be overcome when collecting real patient images.

Deep learning is a promising machine learning approach for analysing medical images. Using deep learning to assess induration images may result in the identification of salient characteristics in the images which could enhance the LTBI screening further. However, training a deep neural network to achieve this goal would require a large number of expertly labelled induration images. These problems motivate the development of a framework for generation of realistic synthetic induration images to increase the size of the induration image dataset. Such synthetically produced induration images could then be used to train a classification network.

1.2 Research Aim and Objectives

As stated earlier, the current mobile phone-based LTBI screening tool consists of a user interface and a backend image analysis system. The aim of this study was to evaluate the backend image analysis framework and explore an alternative method for assessment of the induration images. This was achieved through the following objectives:

1. Evaluation of the image analysis algorithms developed in the previous studies using real clinical induration image data sets collected during this study.
2. Examination of the need for 3D reconstruction during induration analysis by determining whether the depth of the induration or the image orientation influences the diameter measurement.
3. Application of a deep learning generative modelling framework to generate synthetic induration images from a limited induration image dataset. Such images may be used to train a deep neural network to identify salient LTBI characteristics and further enhance the induration image assessment and measurement result.

1.3 Scope and Limitations

Although the existing UCT LTBI screening tool consists of a mobile application and a backend processing system, the focus of this project was to evaluate the backend system, which is responsible for the induration image processing and analysis to obtain an induration diameter measurement, and to explore an alternative. The current system has been developed for use on a low-cost smartphone, and for use in resource-limited areas and this perspective continued in this study. A relatively small number of real images was available for implementation of the objectives.

1.4 Structure of dissertation

The rest of the dissertation is organised as follows. Chapter 2 reviews literature relating to tuberculosis, advances in diagnosis methods for LBTI and relevant mobile-phone based image analysis methods. Chapter 3 gives an overview of the methodology used to meet the objectives of the study. Chapter 4 entails evaluating the LTBI screening tool developed in previous studies using real induration images collected from patients during this study. Chapter 5 investigates the need for 3D reconstruction during the induration diameter measurement procedure in the LTBI screening tool. Chapter 6 discusses the method used to generate synthetic induration images to be used in training a deep learning-based image classification model for latent tuberculosis infection. Chapter 7 concludes the report and discusses limitations and areas for future work.

2 Literature Review

This chapter reviews literature related to the epidemiology and pathology of tuberculosis, as well as existing methods used to detect latent tuberculosis infection. Potential latent tuberculosis screening methods which could be used to improve on the current screening process are highlighted. Literature related to existing alternative screening methods was also reviewed, and the development areas in these new approaches highlighted. Special attention is given to the existing image analysis framework of the mobile phone-based latent tuberculosis infection screening tool currently in development at the University of Cape Town.

2.1 Epidemiology and pathology of the tuberculosis pandemic

According to the World Health Organization (WHO), TB is the leading cause of death from a single infectious disease, worldwide (WHO, 2017). In 2016, 10.4 million people developed active TB, with 1.7 million people dying from the disease (WHO, 2017). Given the fact that effective treatment is available, this has raised significant concern in the global public health sector (WHO, 2017).

The TB disease is caused by the bacillus Mycobacterium tuberculosis (M. tuberculosis) pathogenic bacteria which most commonly attacks and damages a person's lungs (Ellner, 2016). The disease is airborne and is therefore spread by infected patients who expel the M. Tuberculosis bacteria from their lungs into the air through coughing, sneezing or speaking. Consequently, the disease is highly contagious spreading much faster in poorly-resourced and densely populated communities due to limited healthcare services (Glaziou et al., 2015). Over 95% of TB cases and 99% of deaths occur in resource-limited settings (Ellner, 2016).

A third of the world's population, approximately 1.7 billion people, is infected with the M. tuberculosis bacteria but do not have any symptoms of the active TB disease (WHO, 2017). This condition is called latent tuberculosis infection (LTBI). The probability of the active disease developing from LTBI is highly dependent on the health of the individual. Individuals who are infected with human immuno-deficiency virus (HIV), suffer from diabetes, are malnourished or those who are exposed to health risk factors such as smoking or alcohol consumption, are at a much higher risk of developing active TB, compared to a person with a healthy immune system.

Although LTBI is not a compromising infection, it does act as a silo for future active TB. The diagnosis of LTBI allows for those patients who are at the highest risk of the disease progressing to active TB to be treated or vaccinated. Patients can also be made aware of the importance of keeping healthy to prevent the development of active TB. This is a crucial part of tuberculosis control and is necessary to achieve the targets of the WHO End TB Strategy (WHO, 2017).

Mardani and Abtahain (2015) state that one of the greatest challenges in reducing the TB prevalence worldwide is low diagnosis rates of LTBI. Preventative TB medication is widely available, but prior to administration, accurate detection of LTBI is necessary (WHO, 2017). Currently there are only two methods used for LTBI detection; the Mantoux Tuberculin Skin Test (TST) and the Interferon Gamma Release Assay (IGRA) (Mardani & Abtahian, 2015).

Tuberculosis is the most common concurrently occurring disease in HIV infected patients (Manosuthi, Wiboonchutikul & Sungkanuparph, 2016). The human immuno-deficiency virus attacks and weakens a patient's immune system thereby changing the pathogenesis of TB from a slowly progressing to a highly fatal disease (Cahn, et al., 2003). In 2016, one million people living with HIV fell ill with TB, with 0.4 million of these patients dying from the disease (Organization, 2018). Early detection and treatment of LTBI is therefore vital to prevent the disease progressing in HIV positive patients.

2.2 Detecting the presence of LTBI

Several methods exist for detecting the presence of LTBI. These include the TST which involves administration of tuberculin into the patients arm and considering the measurement of the resultant induration diameter together with patient-specific information, and IGRA which is a blood-based test. Emerging methods include mHealth applications which assist in the measurement of the TST induration.

2.2.1 Tuberculin Skin Test

The current, most common method for screening LTBI is the TST (WHO, 2017). The test involves tuberculin being injected into the patient's forearm (Glaziou et al., 2015). A reaction between the M. tuberculosis bacteria, if present in the patient, and the tuberculin protein results in a cutaneous induration at the site of the injection (WHO, 2017). The presence of LTBI is indicated by the diameter measurement of the induration. The standard procedure used when conducting a TST involves

administration of the tuberculin, measurement of the diameter of the resultant induration and interpretation of the results (CDC, 2013).

i. Administration of the tuberculin

The clinician first locates the area approximately 10 cm distal to the patient's elbow joint. Tuberculin is then injected into a patient's arm and forms a wheal under the skin as shown in Figure 2.1.

Figure 2.1: Wheal on patient's arm after injection of tuberculin under the skin (image from CDC (2013)).

ii. Induration measurement

On returning to the clinic, 48 to 72 hours after administration of the tuberculin, the patient's arm is examined for an induration. If present, a clinician marks the edges of the induration using a ballpoint pen, and then measures its' diameter using a ruler. The widest edges of the induration are identified by palpating the induration with fingertips as seen in Figure 2.2a. A ruler is then placed over the induration, perpendicular to the long axis of the arm. The measurement is recorded, by the clinician, to the nearest millimetre as seen in Figure 2.2b. In most patients, a red area develops around the induration. This is known as an erythema, which also occurs in patients who do not develop an induration.

(a) (b)

Figure 2.2: The ballpoint pen (a) and ruler (b) method to identify and measure the induration formed during the TST (image from CDC (2013)).

iii. Interpretation

Diagnosis of LTBI is determined by the size of the induration, the patient's health status and their self-reported living environment. Three induration size measurement ranges are important for determining the diagnosis: ≥ 5mm, ≥ 10mm and ≥ 15mm. Table 2.1 shows the three ranges and the corresponding patient information for a positive LTBI diagnosis.

Table 2.1: Table showing the three size categories of indurations and the corresponding patient-specific information for a positive LTBI diagnosis (CDC, 2013).

Induration Diameter	Condition for a positive LTBI diagnosis.
≥ 5mm	Persons infected with HIV. Recent contact with TB case patient. Persons with indications of prior TB infection. Patients with organ transplants.
≥ 10mm	Recent immigrants from countries with high TB prevalence. Injection drug users. Residents or employees of high-risk congregate settings (i.e. prisons, homeless shelters). Mycobacterial laboratory personnel. Persons with high risk clinical conditions (i.e. silicosis, diabetes mellitus, chronic renal failure). Children under 5 years of age. Infants, children and adolescents exposed to adults at high risk of developing active TB.
≥ 15mm	Persons with no known risk factors for TB.

The TST is both simple to administer and widely available (Mardani & Abtahian, 2015). The simplicity of the test allows it to be performed easily in resource limited areas as it does not depend on a laboratory or any special equipment. The test has a very good sensitivity and specificity, provided patients return for the reading (Linas et al., 2011). In addition, because it was the only diagnostic method for LTBI until the beginning of the century, there has been substantial research done to improve and refine the test as well as validate its application (Mardani & Abtahian, 2015).

However, the accuracy of the test is affected by a number of factors including the depth at which the tuberculin is administered under the skin, the exact quantity of tuberculin which is administered and the crude measurement procedure; the most prominent being human error in the measurement procedure (Meyer, Hougen & Edwards, 1951). During the measurement procedure, a ruler is placed over the induration which may result in deformation, impacting the accuracy of the measured induration diameter. Errors may also arise from the pen markings as their exact positions are left up to the discretion of the clinician (Huebner, Schein & Bass, 1993).

Along with the inaccuracy in measurement, there are other limitations associated with the traditional TST method. The visit to the clinic for the measurement of the induration needs to occur 48 -72 hours after the tuberculin is injected into the patient's arm. This delay is required for the induration, if present, to reach peak diameter before the measurement is done. However, patients often do not return for the follow up reading due to various reasons such as long walking distances to the clinic, forgetting, not having enough time or inability to cover the transportation cost involved in returning to clinic (Risser et al., 1985). In resource limited areas there is also often a lack of skilled clinicians who are able to perform the required measurements and provide a diagnosis (Mardani & Abtahian, 2015). A study by Lina et al. (2011) shows that these factors contribute to unsuccessful reading in 10% of patients. Consequently, this leads to reduced diagnosis capacity, risk of further disease progression in patients, spreading of the disease and wastage of resources.

2.2.2 Interferon Gamma Release Assay (IGRA)

The IGRA test is a blood-based test which determines a patient's LTBI status through measuring their immune response to TB proteins (Mardani & Abtahian, 2015). After blood has been taken from the patient, it is exposed to antigens from the M. tuberculosis bacteria which have been combined with peptides (Pooran et al., 2010). If M. tuberculosis is present in the patient's blood, there is a release of interferon gamma (INF- γ), indicating the presence of LTBI (Trajman et al., 2013).

An advantage of the IGRA test is that results can be obtained within 24 hours of administering the test. Furthermore, there is reduced healthcare worker subjectivity in the assessment as results are determined in a laboratory (Mardani & Abtahian, 2015). An additional advantage is that only a single visit to a healthcare facility is required for gathering all the necessary information and blood samples.

Limitations of the IGRA test include: the need to process the blood sample within 8-30 hours of blood being drawn from the patient, inability to do the test on children under the age of five years, and the test's ineffectiveness for patients who have been exposed to TB prior to testing (Mardani & Abtahian, 2015). Trajman et al. (2013), suggested that replacing TST with IGRA tests in resource limited settings is not recommended due to the higher costs involved in IGRA tests. Furthermore, Linas et al. (2011) and Trajman et al. (2013) agree that the TST method has similar, and in some cases better, sensitivity and specificity than the IGRA test.

2.2.3 mHealth for LTBI diagnosis

mHealth encompasses the use of mobile devices for health interventions and aims to assist in making health services broadly accessible to people across the globe (Latif et al., 2017). Seventy percent of adults above 16 years old in Sub-Saharan Africa have been reported to have a mobile subscription (GSMA, 2017). Africa has an annual mobile penetration growth rate of 65% in rural and remote areas (Stork, Calandro & Gillwald, 2013). Driven by this escalating prevalence of smartphones, mHealth has grown significantly (Albabtain et al., 2014).

A review by Latif et al. (2017) summarizes the different applications of mHealth in developing countries. These include education and awareness, clinical decision support systems, epidemic outbreak surveillance and monitoring, remote diagnostic and treatment support, training healthcare workers and disease management.

In the clinical decision support domain, health insights which have been intelligently filtered may complement clinician and patient knowledge, improving clinical decision-making. Information gained remotely via a mobile phone by clinicians or patients themselves, can be processed to help with challenging diagnostics and prescriptions (Martínez Pérez et al., 2014). This is especially beneficial in areas where healthcare equipment is limited and insufficient staff numbers exist (Albabtain et al., 2014). Mobile technologies have therefore created an opportunity to lower the demand on healthcare facilities and staff and to allow healthcare assessments to be more accessible for patients who have limited access to healthcare services. However, adoption of mHealth applications is still slow in Africa compared to developed countries due to limitations such as low health literacy, language barriers, the high cost of mobile devices, limited internet coverage and limited healthcare infrastructure (Karageorgos et al., 2018). With respect to LTBI diagnosis and in the clinical decision support domain, Naraghi et al. (2018) and Dendere et al. (2017) proposed a mobile phone-based system as an alternative to the traditional measurement procedure in the Mantoux TST method. The purpose of the system is to eliminate the need for patients to return to a clinic for the second measurement visit during the TST process and/or improve the accuracy of the induration measurement. The system acquires and processes images of a patient's induration, promising to remotely obtain the diameter measurement.

There are two potential use cases of the application discussed by Dendere et al. (2017). Firstly, the application could be used by the patients who are being screened for LTBI. In this use case, the patient would need to complete the image capture procedure after the clinic visit and send the images to a remote processing centre themselves. This would eliminate the costs and time associated with a second clinic visit for the patient but with the limitation that the image capture procedure may be difficult to perform without assistance from an experienced user. The second use case involves healthcare workers being the primary users of the application; they may capture images of the TST response in the community. This use case may improve the accuracy of the measurement procedure in areas where healthcare workers have limited experience in reading TST measurements.

The current application requires patients or healthcare workers to send multiple images (a total of 7) of the induration and patient specific information to a remote processing centre, which can lead to excessive data costs for the patient or the health system. At current rates, the cost of buying 100 MB of data from any of the four major service providers in South Africa (Vodacom, Cell C, MTN or Telkom) is R29 (ICASA, 2018). The total size of the packet of data to be sent varies between 8 and 10 MB (Naraghi et al., 2018). This leads to a total cost, incurred by the patient, of approximately R3 for each assessment; for the use case of healthcare workers collecting and transmitting images, the cost to the health system may be lower if agreements are made with service providers. Furthermore, the need for multiple images decreases the usability of the application as a strict image capturing protocol involving images being taken from specific angles and orientations above the induration needs to be adhered to.

Although a mobile phone-based method has the potential for reduced risk of clinician error in measurements (Naraghi et al., 2018), current algorithms used in delineating the induration and acquiring the final diameter are inaccurate. This is because the assessment is done using a two-dimensional (2D) depth map generated from a 3D reconstruction of the induration, making the diagnostic accuracy of the segmentation algorithm highly dependent on the induration height and colouring. Naraghi et al. used Agisoft PhotoScan (APS) to perform the 3D reconstruction of the induration during an assessment. The underlying technique employed in their reconstruction process in this software is structure from motion (SFM); this was successfully employed in reconstructing pronounced synthetic indurations in 3D, which was aided by features such as hair, moles, freckles, and a reference sticker that could act as tie-points across the images. Tie points are essential for the SFM algorithm to work. However, the more subtle indurations were not always present in Naraghi et

al.'s resultant 2D depth map images. This was of concern as most real indurations present as slightly raised welts rather than well-defined bumps.

2.3 Deep learning for image classification

Machine learning models, which provide a statistical representation of training data, may be used in image analysis. Trained models may be used to automate pattern identification utilized in decision-making and extract predictions on future data (Lee et al., 2017).

Deep learning is a form of machine learning which aims to mimic functions of the biological human brain by making use of a multi-layered neural network which allows the transfer of information, through interconnected layers, for analysis and decision-making (Lee et al., 2017). Convolutional neural networks (CNNs) are a subset of deep learning and are most commonly applied to image analysis tasks. They have been successfully used in the field of image classification and segmentation, particularly in medical applications (Lakhani et al., 2018). This is because they enable complex, non-linear functions to be developed which describe the most common features and patterns in images (Yamashita et al., 2018). These functions are then mapped to an output and used as a classification result for the image (Yamashita et al., 2018, Zeiler & Fergus, 2014).

A CNN is comprised of 4 layer types: convolutional layers, which extract the image features in small areas of the image at a time; activation layers, which break the linearity in the model by applying a non-linear function to the features which enables the network to learn complex non-linear features across the data; normalization layers which normalize the features in the model along the dimension of the batch of data; and pooling layers, to which these features are passed. New smaller features are then extracted from the pooling layers and passed through the previously mentioned layers again (Zeiler & Fergus, 2014). The layered structure of a CNN is therefore what allows the extraction of features that are so complex or subtle that they may be unknown or not easily identifiable by humans. This is also what enables certain machine learning algorithms to outperform humans in certain tasks.

The main limitation of deep learning, which has significance for medical applications, is the high dependency on large amounts of high-quality data (Lee et al., 2017). The reason for the lack of large medical image data sets are that medical images are often costly to acquire, and a large amount of work needs to be done by experts to produce and label the images. Furthermore, there are privacy

and ethical issues with collecting and analysing medical images which need to be overcome to ensure that patient integrity is not breached (Lakhani et al., 2018). To overcome the problem of insufficient data sets, the most common and effective methods are data augmentation, transfer learning or the use of generative models such as generative adversarial networks (GANs).

2.3.1 Data Augmentation

A common problem encountered when training a deep learning model on a limited dataset is memorization which occurs when the model becomes too closely fitted to the training data (Wong et al., 2016). For example, instead of grouping medical image data into “affected” and “unaffected”, for a particular disease, an inadequately trained deep learning algorithm would correctly classify the training image dataset and any new image (affected or un-affected) as un-affected as it is not part of the original affected training group. To prevent overfitting or memorization of the training data and increase the accuracy and generalization of CNNs, the training data can be augmented (Wong et al., 2016). Traditional methods of data augmentation include adding noise and applying transformations (rotations, translations, zoom, flipping, shearing and colour perturbation) to images, which are done before the images are fed into the network for training. Many machine learning libraries have built-in data augmentation functions which can be run automatically.

2.3.2 Transfer Learning

Transfer learning utilizes a process suited for one specific task to help in solving a different problem by transferring knowledge from a large dataset, a source domain, to a smaller dataset, the target domain (Lakhani et al., 2018). An example of this would be modifying a deep learning algorithm originally trained on natural images to classify radiographic images. The underlying assumption in transfer learning is that all images are inherently made up of similar features such as edges and blobs enabling algorithms to be manipulated for a variety of different applications (Tajbakhsh et al., 2016).

Shin et al. (2016) demonstrated that pre-trained networks often perform better than models trained from scratch, regardless of training data size. The authors reported thoraco-abdominal lymph node detection and interstitial lung disease classification by fine-tuning CNN models pre-trained from natural image datasets. They concluded, in agreement with Zhu et al. (2012), that when searching for an optimal solution, emphasis should be placed on considering the trade-off between using better transfer learning models as opposed to using more training data.

Lakhani et al. (2018) successfully used transfer learning to develop a classifier that differentiates between chest and abdominal radiographs using only 65 training images. This was done by removing the final fully connected layers of the pre-trained model and inserting additional layers with random initializations, to allow the model to learn from the new medical data.

2.3.3 Generative Adversarial Networks

Generative adversarial networks (GANs) provide an alternative to data augmentation or transfer learning by producing realistic synthetic images to assist in training a model. They consist of two neural networks, one acting as a generator, which generates synthetic images, and the other as a discriminator, which tries to classify the generated images as synthetic compared to the training images (Goodfellow et al., 2014). These networks contest with each other until the images produced by the generator are indistinguishable from real images, from the discriminators point of view.

GANs have been successfully used for reproducing realistic images of skin lesions, indistinguishable from real ones as determined by expert dermatologists (Baur, Albarqouni & Navab, 2018). Baur, Albarqouni and Navab (2018) used the progressive growing of GANs (PGAN) in their work as it has shown outstanding results when generating high resolution images, up to 1024 x 1024 pixels (Karras et al., 2018). The PGAN starts by generating low resolution images and new layers are then added that model increasingly fine details in the images during the training progresses (Karras et al., 2018). Baur, Albarqouni and Navab (2018) assessed the realism of their generated images both qualitatively, by performing a user study involving 3 expert dermatologists as well 5 deep learning experts, and quantitatively using the sliced Wasserstein distance (SWD) which compares the image distributions of the real and generated images (Karras et al., 2018). The qualitative evaluation resulted in a classification accuracy of slightly above 50%, implying that the experts are only slightly better at classifying the images when compared to a completely random classification of the images. This shows that, to the human eye, the generated images resemble the real images.

An advantage of using synthetic images for training a neural network is that there are no privacy or ethical issues which need to be addressed when using these images, which is beneficial in medical applications where security of the data and images is sensitive (Lakhani et al., 2018).

2.3.4 Development toolkits for deep learning

Erickson et al. (2017) and Wang, Zhaobin et al. (2019) describe the myriad frameworks available for machine learning research and application development. The most actively used frameworks for medical imaging tasks are TensorFlow (Google: https://www.tensorflow.org/) and Pytorch (Google: https://pytorch.org/) (Lakhani et al., 2018). Table 2.2 highlights the differences between these two frameworks.

Table 2.2: Comparison between Tensorflow and Pytorch frameworks

	Tensorflow	**Pytorch**
Developed by:	Google	Facebook
Learning curve	Steep learning curve.	Easier to learn as it is pythonic and intuitive.
Community	Large community with many resources and support pages.	Small community with less content than Tensorflow.
Additional Tools	Tensorboard for visualization of models in the browser.	No additional tools for visualization of images
Graphs	Static graphs where the entire computational graph needs to be defined before running the model.	Dynamic graphs which may be defined and manipulated on-the-go. This is useful when using variable length inputs.

Tensorflow is regarded as a more suitable framework for development of systems to be used in production and large-scale projects while Pytorch dominates in research applications as it is more intuitive to use and works well for rapid prototyping (Lakhani et al., 2018).

Google Colaboratory (Google: https://colab.research.google.com) is a convenient and free alternative coding environment for either buying a machine with the required computing capabilities for deep learning or renting a more powerful cloud computing machine at a cost. The tool allows the computing to be off-loaded from the local machine onto the Colaboratory machines located in the cloud, where it is able to train continuously regardless of whether the local machine is still running or not. A limitation with Colaboratory is that the training data need to be constantly available to the machines on which the code is being run. This can be resolved by allowing the machine to mount a google drive folder where the data are permanently stored. This also allows for the generated images to be saved to an accessible location on the same drive.

2.3.5 Evaluation metrics for synthetic images

The generator model of a GAN has no objective measure on which it can be evaluated since it is trained by a second discriminator model. Therefore, the performance of a GAN needs to be evaluated on the quality of the synthetic images that it generates (Salimans et al., 2016). Although there is no consensus around the best quantitative evaluation methods for GANs, the two most widely adopted methods are the inception score (IS) and the Fréchet inception distance (FID) score (Borji, 2019).

The IS, proposed by Salimans et al. (2016), uses a pre-trained image classification model to analyse the synthetic images. It does this by determining the probability of the synthetic images belonging to each of the image classes in the model. The probabilities of each synthetic image belonging to each of the image classes are then aggregated and evaluated against the other synthetic images to determine the inception score of the GAN. A drawback of the IS comes from the fact that the statistics of the target images are not compared to the statistics of the generated images. This may result in the generated images scoring well in terms of IS if they are realistic looking images, but still not adequately representing the target domain well (Heusel et al., 2017).

The FID was proposed by Heusel et al. (2017) as an improvement on the inception score. The FID score models the distributions of the generated and real image sets and measures the distance between the two distributions, as opposed to the IS which only measures the diversity and quality of the generated images (Borji, 2019). The Fréchet distance, which is used in calculating the FID score, is often explained using the analogy of a man walking a dog on a leash. If we track each of their paths, we are left with two curves. The Fréchet distance is the minimum leash length required to allow each to follow their own path. The shorter the leach length, the greater the similarity between the man's and the dog's paths. Similarly, the FID score decreases as the distributions of the two image groups being compared become more alike (Heusel et al., 2017). A low FID is therefore desirable as it represents a group of images that closely mimic the dataset of real images. Because the goal of the FID score is to capture the accuracy of the model in generating synthetic images which specifically represent the target images, it performs particularly well when analysing the synthetic images from a GAN trained to generate a very specific class of images (Heusel et al., 2017).

Qualitative assessments, on the other hand, are necessary during the training phase of GAN generation as human intuition helps greatly with inspection and tuning of models. However, these assessments are not sufficient to evaluate the accuracy of the model. This is because visual assessments are subjective in nature and could be biased to models that overfit as the images seem realistic but the model may not produce the required variance in the images (Borji, 2019). Furthermore, qualitative assessments are cumbersome, difficult to reproduce and often require adequate knowledge of the target domain (Borji, 2019).

2.4 Summary of literature review

The Mantoux TST method is the most popular method used to detect the presence of LTBI in resource-limited areas due to its low cost, reliable performance and simplicity of use. However, limitations such as the need for a second clinical visit by patients, has motivated research into a mobile phone-based tool for LTBI screening. High rates of mobile penetration in rural, remote and resource limited areas make the mobile-phone based LTBI screening tool a viable solution.

Although the prototype LTBI screening tool (Naraghi et al., 2018) addresses the limitations of the Mantoux TST, the algorithms that have been developed for use in delineating the induration and acquire the final diameter measurement have only been tested on simulated indurations. Hence, there is a need for the system to be tested on real induration images from patients. Furthermore, deep learning could be explored for image classification and identification of salient characteristics to enhance diagnoses, but with the limitation that it requires a large dataset of induration images for training. Large image datasets are often scarce in the medical field. Generative adversarial networks provide an opportunity to overcome this limitation as they can produce synthetic images which are similar to the original images. However, to date, there have been no reports in the literature on the generation of synthetic indurations.

3 Methodology Overview

The aim of this study was to evaluate the backend image analysis framework of the LTBI screening tool and explore an alternative approach to induration image assessment through deep learning. The following methodology was used to meet the above-mentioned aim (see Figure 3.1 for an overview).

First, Chapter 4, describes evaluation of the image analysis component of the earlier LTBI screening tool using a dataset of real induration images. This encompassed the collection of real induration images and evaluation of the effectiveness of the image analysis algorithms in analysing the collected real induration images.

Second, evaluation of the need for the 3D reconstruction of the induration is described in Chapter 5. This was done by determining whether the depth of the induration or the image orientation influenced the resultant diameter measurement during the segmentation process. The results highlighted that the 3D reconstruction aspect was not necessary for the assessment.

Chapter 6 describes the design of a system to produce synthetic induration images to increase the size of the available induration image dataset for later use in training an image classification model. A robust GAN was adapted first to ensure realistic image generation using a large dataset of melanoma images. Then, using this architecture, the network's parameters were manipulated to allow realistic image generation from a small subset of the melanoma image dataset. Once the network could produce realistic melanoma images from the limited dataset of 200 images, the same network architecture and parameters were then used to produce synthetic induration images from a dataset of 150 induration images.

Ethics clearance for this study was granted by the human research ethics committee (HREC) at the University of Cape Town (HREC 319/2018).

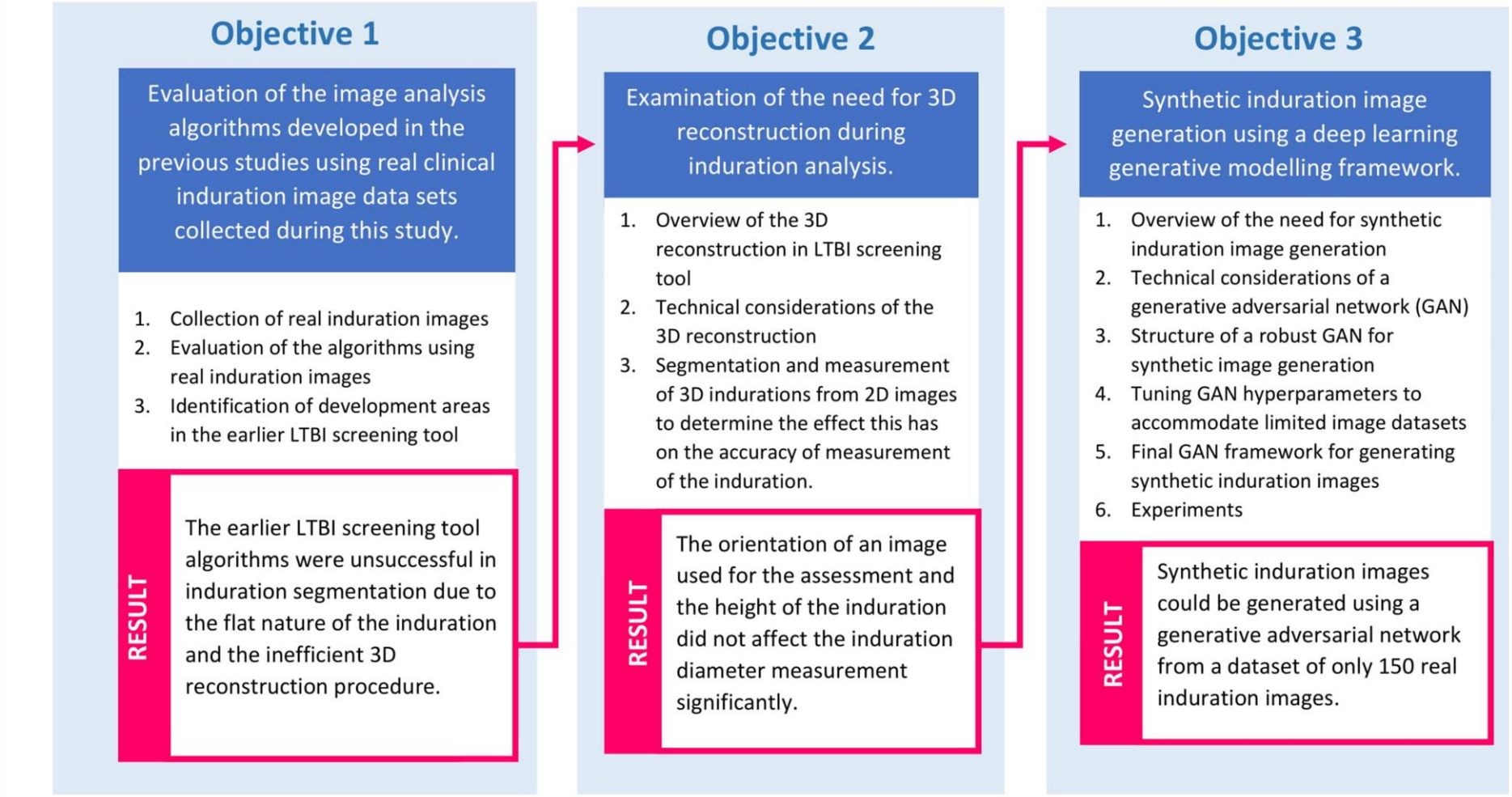


Figure 3.1: The approaches used to achieve the objectives and overall aim of the project

4 Evaluation of the LTBI screening tool using real induration images

This chapter describes the evaluation of the latent tuberculosis infection (LTBI) screening tool developed in previous studies using real induration images collected from patients during a Tuberculin skin test (TST) screening exercise which was being conducted in Cape Town. The first section gives an overview of the LTBI screening tool. The second section describes the technical considerations for the tool. Modifications to the software packages used in the LTBI screening tool are documented in the third section. Following this, the methodology used for image acquisition of the real induration images from patients is described. Evaluation of the effectiveness of the image analysis algorithms developed in previous studies in analysing the collected real induration images, is presented next. Finally, the results and conclusions drawn from the evaluation of the screening tool using real induration images are discussed.

4.1 Overview of the LTBI screening tool

The LTBI screening tool proposed by (Naraghi et al., 2018) and (Dendere et al., 2017) is a mobile phone-based application which requires patients to capture images of the resultant induration during a TST to allow processing of the test results at a remote processing centre. Prior to the tool being used, the patient is required to complete the initial phase of the TST (administration of tuberculin to the patient's arm) at a healthcare facility. When the induration has reached its peak diameter, 48-72 hours after administration of the tuberculin, the LTBI screening tool may be used as an alternative assessment method to the patient returning to the healthcare facility for induration measurement. The patient is required to follow the prompts in the application interface which guide the patient into capturing relevant information for the test as well as 7 images of their induration. During the image capturing procedure there are positioning guidelines and prompts in the application interface which ensure that the images are captured from the correct position above the induration. Once the information and images are captured, the application prompts the patient to send them in the form of an email to the remote processing centre. Once received at the processing centre, the diameter is measured by the software. This involves producing a three-dimensional (3D) reconstruction of the induration from the attained images, extracting a two-dimensional (2D) depth map from the reconstruction and finally segmenting the depth map image to obtain the induration diameter.

4.2 Technical considerations of the LTBI screening tool

The existing mobile application was designed with specific image capturing protocol in place. This ensures the images are captured following the correct procedure to allow successful processing at the remote processing centre. The following protocol must all be adhered to during the image capturing procedure (Naraghi et al., 2018):

- The scaling marker needs to be present in the images to provide a size reference in the images.
- The scaling marker needs to be placed perpendicularly to the long axis of the arm.
- Seven images of the induration site from different perspectives with at least 50% overlap, must be acquired.

Embedded onscreen guidance components ensure the guidelines above are strictly adhered to. Figure 4.1 shows the current application interface with the guidance components as well as the scaling sticker, as designed by Naraghi (2018). The blue arrow indicates the direction of the arm axis and the yellow and green marker boxes must be positioned to encapsulate the scaling sticker and the induration, respectively. The green text at the top of the screen shows the pitch and roll with the green arrow (which changes to a red cross when the orientation is incorrect) guiding the user to capture the images at the correct orientation.

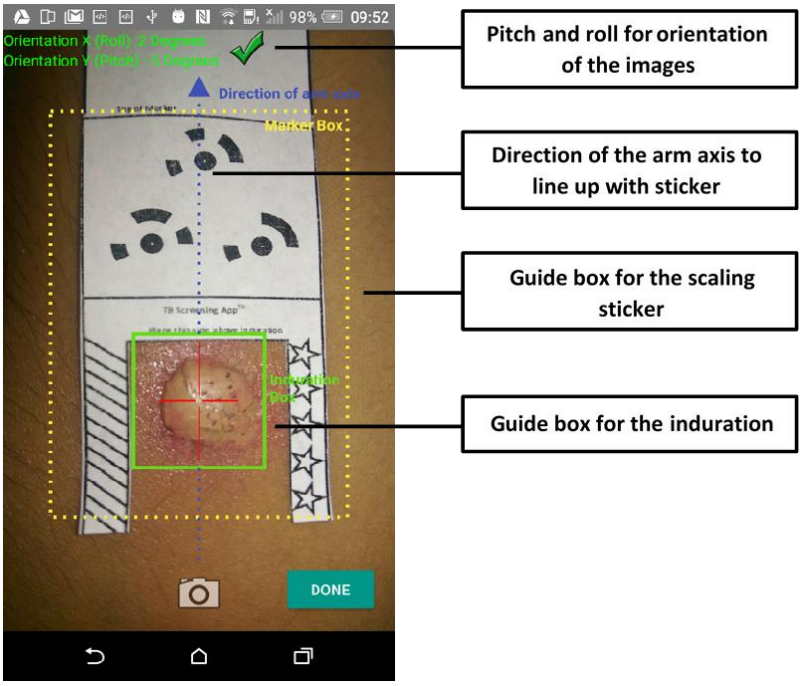


Figure 4.1: The scaling sticker and application interface of the previous LTBI screening tool with guidance components to help the user capture images of the induration from the correct position above the arm (image adapted from Naraghi (2018)).

Once the induration images have been received at the remote processing centre the first phase of the induration analysis is the 3D reconstruction procedure. This procedure is performed using the Agisoft Photoscan (APS) software. Agisoft Photoscan's 3D reconstruction method is rooted in the process of structure from motion (SFM) and is made up of the following four processes: importing the images into the APS environment, aligning the images, generating a dense point cloud and fitting a mesh over the points to produce the 3D reconstruction. These processes are detailed in the succeeding chapter. The two most important considerations being that the images used for reconstruction need to be of high quality and there needs to be an overlap of at least 50% between consecutive images.

Once the 3D model of the induration is available, a 2D depth map is extracted. The 2D depth map represents the change in depth of the surface of the induration and surrounding skin using grey-scale intensity. Using the resultant 2D depth map, segmentation of the induration is done which allows for the diameter measurement to be obtained.

The following steps are taken in segmenting the induration and measuring its diameter:

Step 1 - The 2D depth map image is cropped. This is done by mapping the pixel coordinates of the green induration box, which appears on the capture screen as seen in Figure 4, to the final depth map image using the markers on the scaling sticker to provide a reference. The outer parts of the image are cropped away, leaving only the induration box.

Step 2 - Histogram equalisation is then applied to the cropped image to enhance the contrast in the image (Kim, 1997). This technique enhances the intensity of pixels by increasing the range over which the pixel intensity values fall and results in a greater contrast in the image. This improves the performance of the segmentation method as the contour of the induration is exaggerated through this process.

Step 3 - Bimodal segmentation of the induration is then done using a variation of Otsu's method for thresholding (Otsu, 1979). Otsu's original method considers the bi-model histogram distribution of the pixel intensities in an image and divides the image into two-pixel classes which represent

foreground and background. The threshold value is found from the point where the minimal intra-class variance occurs as seen at point *t* in the example image distribution in Figure 4.2.

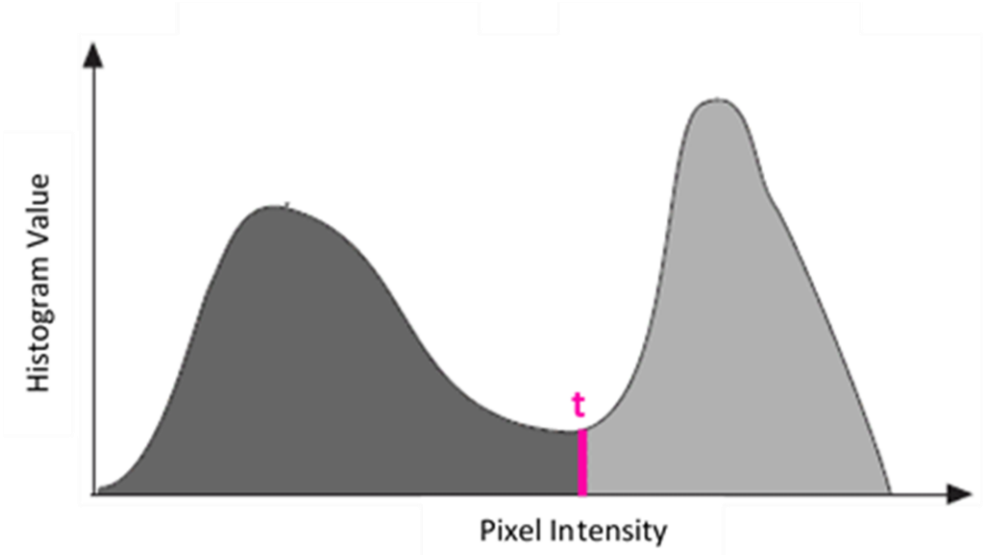


Figure 4.2: Histogram showing the pixel intensity distribution of an image with the minimal intra-class variance value dividing the pixel intensities into two-pixel classes and shown in pink at point t as the threshold value.

The variation of Otsu's method, as presented by Moghaddam & Cheriet (2012), is used in Naraghi et al.'s work (2018). The induration image is divided into localised areas and each area is analysed individually to find the appropriate threshold value for that area. Figure 4.3 shows a depth map image of an induration which has been divided into 8 sections denoted by areas A1 to A8. A threshold value is established for each segment using Otsu's method and is applied to the respective area. This

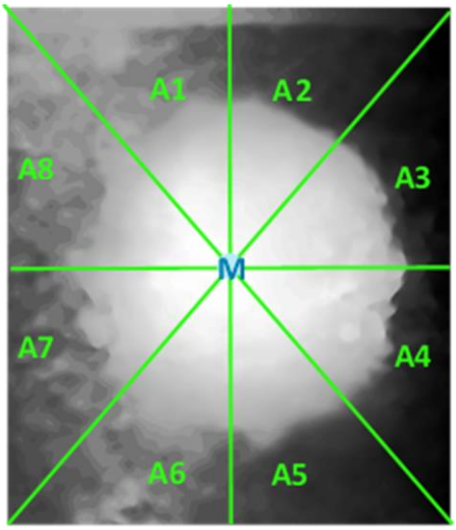


Figure 4.3: Otsu's modified segmentation method where 8 image segments (represented by areas A1 to A8) are each analysed individually.

method allows improved segmentation to be established in images where the contrast between skin and induration is not homogeneous across the whole image.

Step 4 - The contour of the induration is then found using a rough margin identification procedure. This is performed in three steps. Firstly, the edges of features in the image are found using an edge detection algorithm from the OpenCV library (www.openCV.org). A contour detection algorithm from the same OpenCV library is then used to find the largest contour represented by the found edges. This contour corresponds to the boarder of the induration and is finally highlighted using green pixels.

Step 5 - To complete the induration measurement procedure, the green pixels, which signify the edge of the induration, are found and stored in an array. Using the Python EllipseFitter package, an approximated ellipse is then fitted to this contour using the values in the array. The final diameter measurement is then extracted from the diameter of the ellipse as shown by the blue arrow in Figure 4.4.

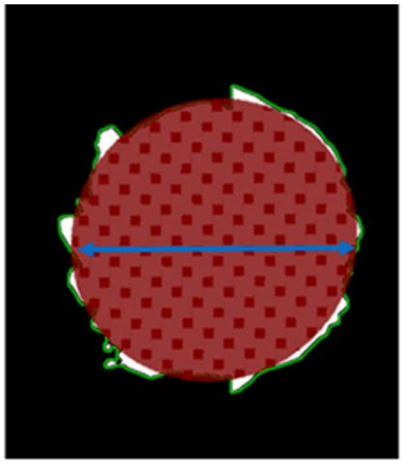


Figure 4.4: A red ellipse fitted to the green induration contour from which the induration diameter measurement can be extracted. The blue arrow highlights the diameter measurement.

The programs and packages used by the existing system for image analysis included Agisoft PhotoScan (http://www.agisoft.com/), a tool for photogrammetry; Python (https://www.python.org/), the programming language used in developing the application; Spyder (https://www.spyder-ide.org/), an integrated development environment for programming in the python language; and the following python packages: Pyautogui, OpenCV2, Mahotas, os, Numpy, Pylab, PIL, Xlutils, EllipseFitter.

The previous version of the mobile application code was written in Android Studio (https://developer.android.com/studio). In this study, the code was modified to allow capturing of 9 induration images rather than 7. This allowed the algorithms to be assessed with added redundancy.

4.3 Collection of real induration images from patients

Images were collected from patients who were participating in a parallel TST study (HREC 702/2017) which was being conducted by researchers from the University of Stellenbosch. Ethics clearance to acquire the real images was obtained before commencement of data collection and informed consent to participate in the study was also obtained from the participants.

Two visits were made to the respective clinics for each patient. The first visit involved obtaining consent from the participant to participate in the data collection process. The clinicians from the HREC 702/2017 study also performed all the relevant clinical procedures and administered the tuberculin during this visit. The second meeting with the participants was made 72 hours after administration of the tuberculin and involved capturing images of the resultant tuberculin reaction by researchers in this study.

4.3.1 Image acquisition

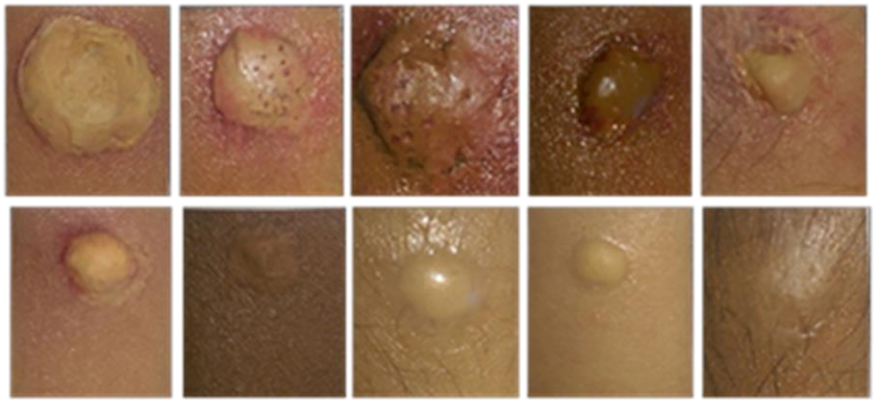
Images were acquired before the traditional ruler and pen diameter measurement. This order would prevent the pen markings made during the measurement from appearing in the images and potentially influencing the image analysis. Furthermore, the images were captured using the existing mobile application protocol as specified by Naraghi et al. (2018) to ensure consistency of image capture. Once captured, the images were sent via email and downloaded to a password protected folder on the author's desktop for further assessment and analysis.

4.3.2 Ground truth diameter measurement acquisition

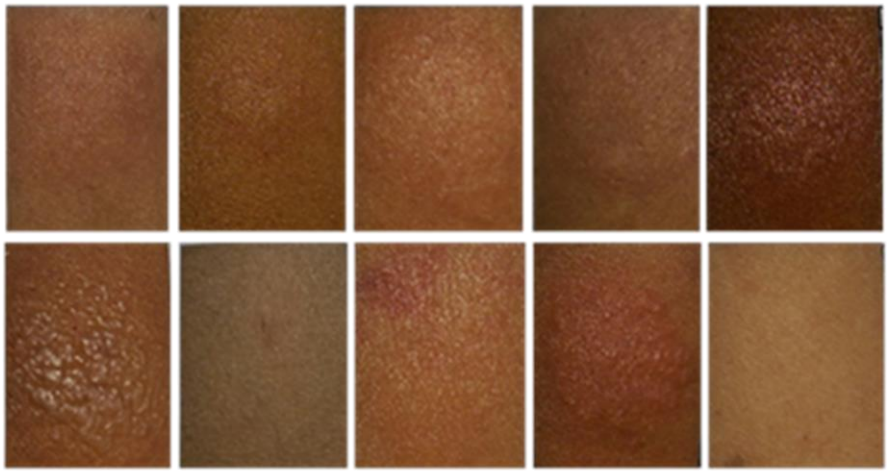
The ground truth diameter measurements were captured by clinicians from the parallel tuberculosis study (HREC 702/2017 study), using the traditional ball-point pen and ruler measurement method. This is the current gold standard measurement method for TST reactions.

4.4 Running the LTBI screening tool algorithms using real induration images

The real induration images differed from the mock induration images which the earlier LTBI screening tool was tested on, as seen in Figure 4.5. The real indurations were much flatter and less conspicuous compared to the mock induration images.



Mock Induration Images



Real Induration Images

Figure 4.5: (a) Sample of mock induration images designed by a makeup artist. (b) sample of real induration images taken using smart phone camera during the TST data collection.

The modified code was run using 10 sets of real induration images. The resultant depth maps for each induration are shown in Figure 4.6.

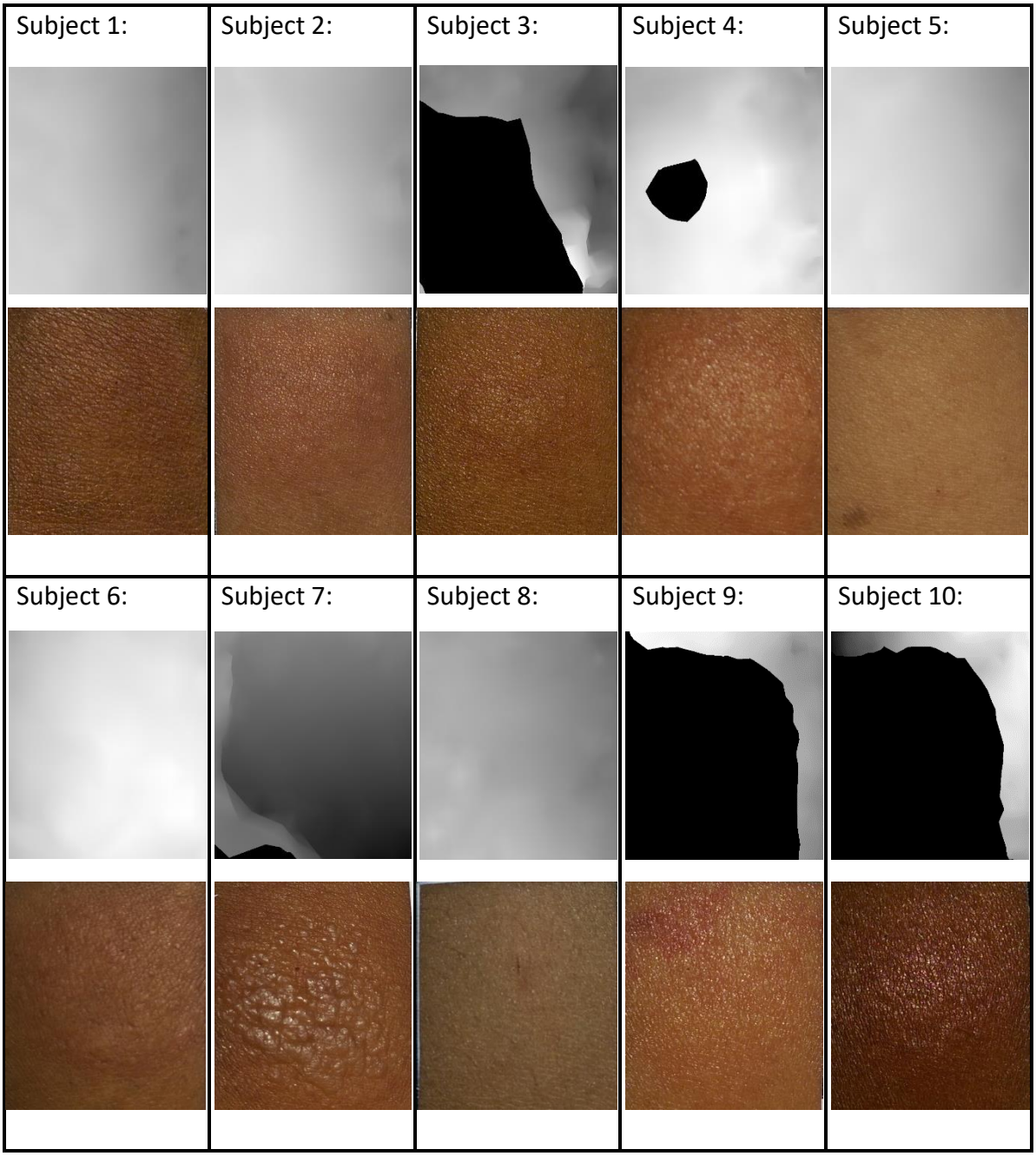


Figure 4.6:Original induration images with resulting depth maps after analysis using existing system.

4.5 Identification of development areas in the earlier LTBI screening tool

The analysis of the induration diameter in the earlier screening tool was done exclusively from the 2D depth map images produced by the 3D reconstruction module. The success of the analysis was therefore dependent on the clarity of the induration in the depth map image. Furthermore, the clarity of the induration in the 2D depth map image was shown to be highly dependent on the height of the induration above the skin surrounding it. It is evident from the images in Figure 4.5 that the height of each real induration was much less prominent, and almost negligible, compared to that of the mock

indurations. This observation is shown in the resultant 2D depth map images shown in Figure 4.6 as there is no trace of the indurations being highlighted or even present in their respective 2D depth map images. This indicates that the depth of the induration is not prominent enough to be identified and shown in the depth map images. Further image processing would therefore be required to highlight the induration more clearly for the succeeding delineation processes to be successful.

Additionally, there are black holes in the final 2D depth map images of subjects 3, 4, 7, 9 and 10. As specified by the APS user manual, these holes may arise as a result of the depth filter removing whole clusters of pixels from the 3D model as they are seen as noise. This is a result of a depth filter being too aggressive or when the reconstruction of the object is done using few images or images of low quality. In this application, the least aggressive depth filter possible was used and the resultant depth map images still contained holes, highlighting the inefficiency in using few images or low-quality images in the assessment. This suggests that quality images should be explored for the LTBI screening tool. However, this is an unfeasible solution since the screening tool should use as few images of as low quality as possible for the assessment to reduce the data costs involved in sending the images from the mobile device to the remote processing centre and to increase usability of the application. If the image processing algorithms in the screening tool are to remain similar, research needs to be done into ensuring low costs for the user without having to compromise on image quality that could negatively affect the test results.

4.6 Discussion

In this chapter, it has been shown that the 3D reconstruction aspect of the LTBI screening tool does not produce the results desired to give an accurate measurement of the induration when real induration images are used. The quality of the 3D reconstruction of the induration and that of the resultant 2D depth map image is vital for the success of the image-based LTBI screening tool. It was found that real indurations are not prominent enough to give clear 2D depth map images as their height is almost negligible. Their flat and inconspicuous nature renders the segmentation aspect of the 3D assessment ineffective. Furthermore, the 3D reconstruction procedure in the screening tool requires more images of higher quality than are currently being provided to perform a successful reconstruction. This defies the need for the screening tool to be low cost and easy to use. As indurations could not be clearly identified in the depth map images, the succeeding segmentation steps in the induration delineation procedure, which aim to find the diameter measurement, could

not be completed. Consequently, it was not possible to obtain TST results from the current LTBI screening tool from real induration images that could be compared to the ground truth measurements recorded by the clinicians from the HREC 702/2017 study.

These findings motivated the work presented in the succeeding chapter, to address the second objective of this study, which sought to establish whether the depth of the induration or the angle at which the induration images are taken affects the resulting diameter measurement and ultimately whether 3D reconstruction is necessary in the assessment.

5 Assessing the need for 3D reconstruction of the induration

This chapter investigates the need for three-dimensional (3D) reconstruction for induration diameter measurement in the latent tuberculosis infection (LTBI) screening tool. This is the second objective of this study. The feasibility of analysing two-dimensional (2D) images for the measurement of the induration is therefore explored.

The purpose of the 3D reconstruction process in the earlier LTBI screening tool was to highlight the induration during image analysis to aid in the delineation process and to normalise the image angle and orientation from which the induration is assessed. As discussed in Chapter 4, the methodology used in the previous LTBI screening tool involved extracting a 2D depth map image from the 3D reconstruction of an induration to be used for further analysis of the induration diameter. The induration diameter analysis, used as an indication of the presence of LTBI, is done exclusively from the 2D depth map image. The results from Chapter 4 showed that the 2D depth map images did not highlight the induration for delineation. This is due to the induration not being prominent enough to be identified in the depth map images and is therefore not evident in them. Hence, research was done into the effect of the induration image position or the induration height on the accuracy of induration diameter measurement when determined from a single 2D image of the induration.

5.1 Technical considerations of the 3D reconstruction module in the LTBI screening tool

The 3D reconstruction procedure used by the LTBI screening tool was developed by Naraghi (2018) using the Agisoft Photoscan (APS) software as discussed in the previous chapter. All processes in the 3D reconstruction process (outlined below) can be automated in APS, through Python batch scripts.

i. Importing the images into the APS environment

The images are retrieved from a specific folder on the computer as specified by the Python batch processing script and imported into the APS environment.

ii. Tie-point detection and image alignment

Agisoft Photoscan scans the images to find common points, known as tie-points. The tie-points are matched across the images and are used to determine the positions and orientations of the images in the 3D space around the object. In Figure 5.1, the blue shapes represent the image positions and orientations in the 3D space. This is shown from two different perspectives to highlight the 3D nature of their positions in the space. Having a scaling sticker with distinct features and markers, as required in the earlier LTBI image capture protocol, improves the calibration of these image parameters by increasing the number of tie-points used to determine the image positions and orientations.

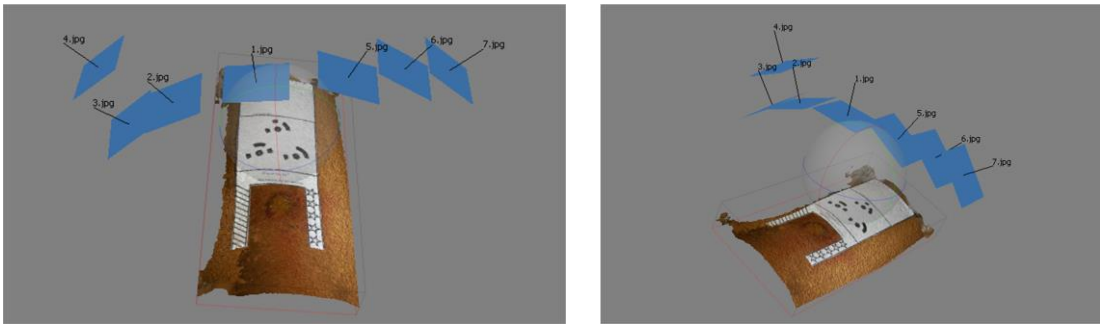


Figure 5.1:Initial 2D image positions and orientations determined by common tie-points in APS and shown as blue shapes in the 3D space around the object.

iii. Dense cloud generation

The estimated image positions and orientations are used to build a dense point cloud. This is done by estimating the depth information in each image (the distance of the object's surfaces from the viewpoint) and producing corresponding depth maps. The depth maps are then projected, from their estimated positions in the 3D space, to a common area where the overlap is used to form the 3D model. These depth maps are passed through a filtering algorithm to remove any points which appear as outliers (points which differ significantly from their surrounding points) to eliminate noise from appearing in the final model. When using a small dataset of images (<10 images) or images of low quality (<5 MB) an aggressive depth filter may eliminate clusters of the dense point cloud which form part of the object. This results in holes on the object's surface. Thus, according to the APS user manual, a moderate depth filter should be used in this application as there are few images (9 images) and they are of low quality (<5 MB).

iv. Mesh generation and surface estimation

To form the 3D mesh and surface of the object, APS interpolates between points in the dense point cloud using polygons to form the surface. The polygon count is set as a parameter and corresponds to the maximum number of polygons used to represent the mesh. The polygon count needs to correspond to the level of surface detail required in the reconstruction. In this application the polygon count is set to high, as a high level of detail is required in the 3D reconstruction.

From the final 3D model, a 2D depth map of the induration box is then extracted for the automated induration delineation process. Usually the depth map is generated from a set distance above the *x-y* plane and is manually input into APS. However, APS has the functionality to use markers detected on the point cloud as a reference for the depth map scale, which eliminates the need for human intervention. In this analysis, three markers on the scaling sticker were used as a reference to provide scale to the 3D object. The depth map is then generated from the top view of the object at the specified height, parallel to the *x-y* plane of the model, to ensure that the subsequent image analysis is done from a normalized image position and orientation. As shown in Figure 5.2 the height of every point in the 3D mesh from the *x-y* plane (corresponding to the height of the surface of the skin) is represented by a *z*-coordinate value. These *z*-coordinate values are used to create the depth map by having each *z*-value correspond to a grey-scale intensity. The minimum and maximum *z*-values are mapped to 0 and 255, respectively; all the other values are transformed and mapped to values between 0 and 255.

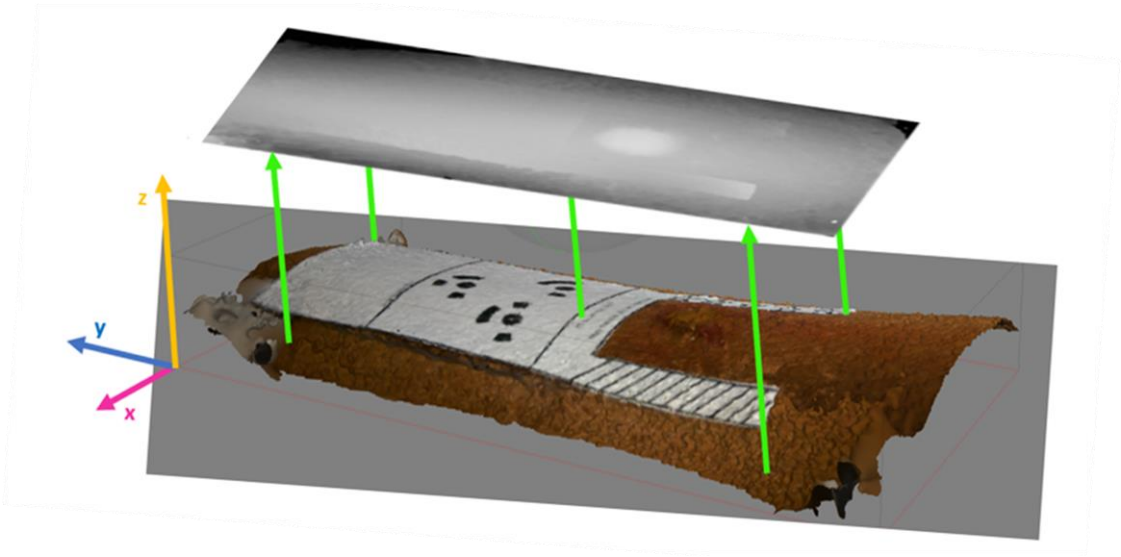


Figure 5.2: The height of every point on the 3D mesh surface, from the x-y plane, (represented by a z-coordinate value and green arrows) is projected onto a plane using the z-coordinate values and their corresponding grey-scale intensity to show the height variation in the image.

The 3D reconstruction procedure therefore relies on the height of the object being assessed and the accuracy relies on the quality and number of images used in the assessment. The flat and inconspicuous nature of the real indurations renders the assessment inefficient and creates the need for many high-quality images for the assessment. This contradicts the need for the screening tool to be low cost and easy to use. Thus, the following sections involve the segmentation and diameter measurement of 3D domes from 2D images taken using the screening tool. This was done to determine the effect on the accuracy of a diameter measurement of a 3D object when using a 2D image for the assessment. It further motivates whether 3D reconstruction is necessary in the evaluation of the diameter of real indurations.

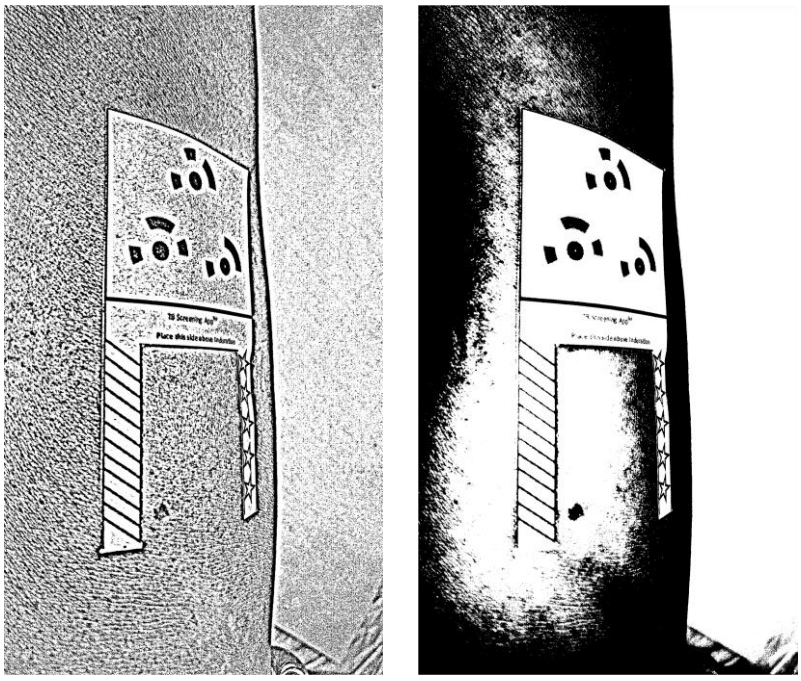
5.2 Segmentation and diameter measurement of 3D domes from 2D images to determine the effect this has on the accuracy of diameter measurement of the domes.

To test whether the angle and orientation of capture of the images using the screening tool affects the resultant dome diameter measurement, the induration box in each image needed to be identified, cropped and normalised before the dome could be segmented and assessed to determine its diameter.

The following steps were taken to extract the induration box from the original images, using functions from the OpenCV2 python package (OpenCV, Google: https://opencv.org/), were:

a) Highlighting the border of the scaling sticker

   Thresholding functions in image processing convert a 2D input image to a binary image by converting each pixel to either black or white based on whether the pixel's original intensity is above or below a set threshold value. Due to the varying illumination of areas in the induration images, a global thresholding function (where one threshold value is set for all pixels in the image) was not effective. An adaptive threshold function, which determines a unique threshold value for each pixel in an image, based on a small region around the pixel, was effective in showing all the borders in the induration image regardless of varying illumination in the image and was therefore used. The cv2.adaptiveThreshold function from the OpenCV library was therefore used to highlight the boarder of the scaling sticker in the images. Figure 5.3 illustrates the results of the two thresholding methods being applied on the same induration image.



Adaptive Thresholding Global Thresholding

Figure 5.3: Adaptive and Global Thresholding applied on an induration image.

b) Fitting a contour line to the boarder of the scaling sticker.

The cv2.findContours function was applied to the image to find all the contours in the image. This algorithm finds contours in an image by identifying pixels of similar intensity at the boundaries of shapes and storing the coordinates of the pixels for each contour in an array. To find contour of the scaling sticker (which is the largest shape in the image), the area enclosed by each contour was calculated using the cv2.contourArea function and the contour corresponding to the largest shape found.

c) Identifying the 4 vertices of the scaling sticker

The cv2.approxPolyDP function was used to find the 4 vertices of the polygon shaped scaling sticker, from its contour. The function is an implementation of the Douglas-Peucker algorithm (Douglas & Peucker, 1973) which simplifies a given shape to an approximated version of the shape with a lower number of vertices, depending on the specified precision. It does this by eliminating points on a given curve (composed of line segments) which are within a set distance from an approximated curve. This is shown in Figure 5.4 where all the red points are eliminated, and the green points remain forming the new vertices of the polygon.

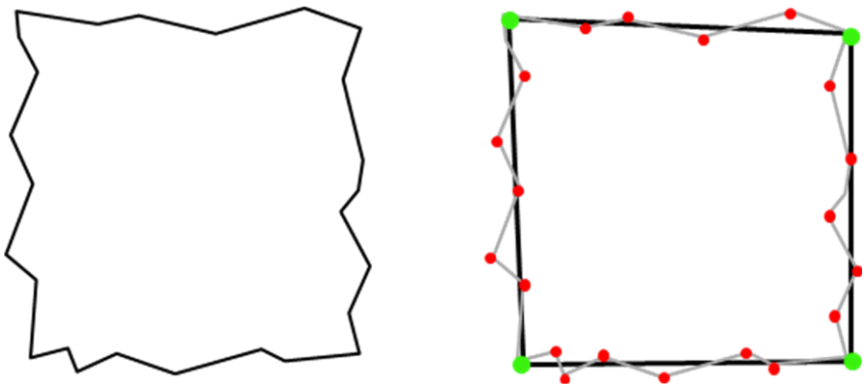


Figure 5.4: Polygon approximation using the Douglas-Peucker algorithm.

d) Perspective transformation of the scaling sticker.

A linear transformation involves mapping all the points in a given quadrangle to another quadrangle and is implemented by multiplying each point in the quadrangle by a matrix. There are two types of linear transformations – perspective and affine. Perspective transformation functions do not preserve parallelism during the transformation whereas the affine transformation does. This concept is illustrated in Figure 5.5. As can be seen in Figure 14, the affine transformation has distorted the square while keeping all the previously parallel lines parallel, but the square which has undergone the perspective transformation has no parallel lines remining after the transformation. The perspective error in the induration images needs to be corrected to normalise the shape and size of the resultant induration box images for consistent assessment. Therefore, for this application, the perspective transformation functions (cv2.getPerspectiveTransform and cv2.warpPerspective) were chosen over the affine transformation functions.

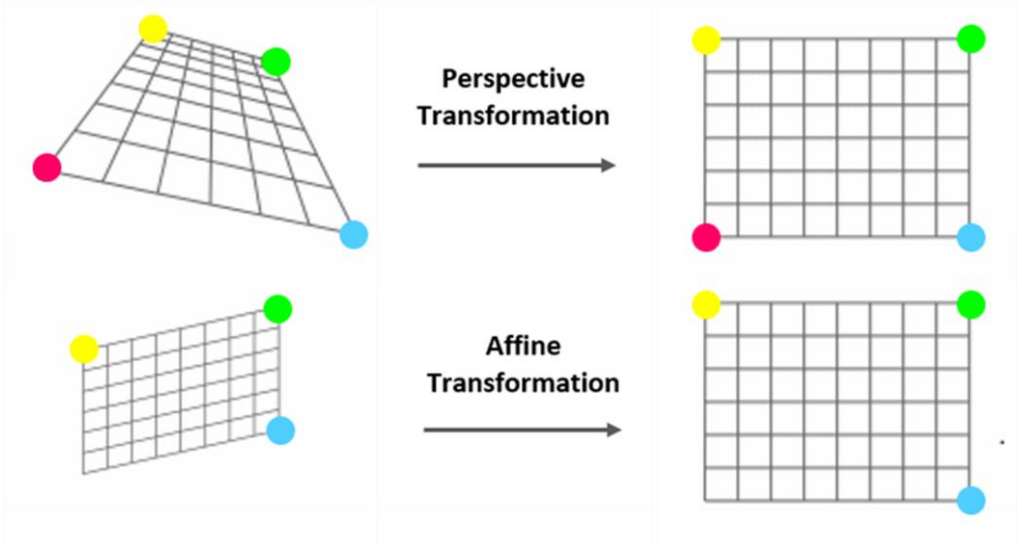


Figure 5.5: Comparison between perspective and affine transformations.

Once the transformation is mapped out, the cv2.resize function is used to transform the scaling sticker to a normalised shape and size. Figure 5.6 shows an original image with the corresponding isolated and transformed scaling sticker.

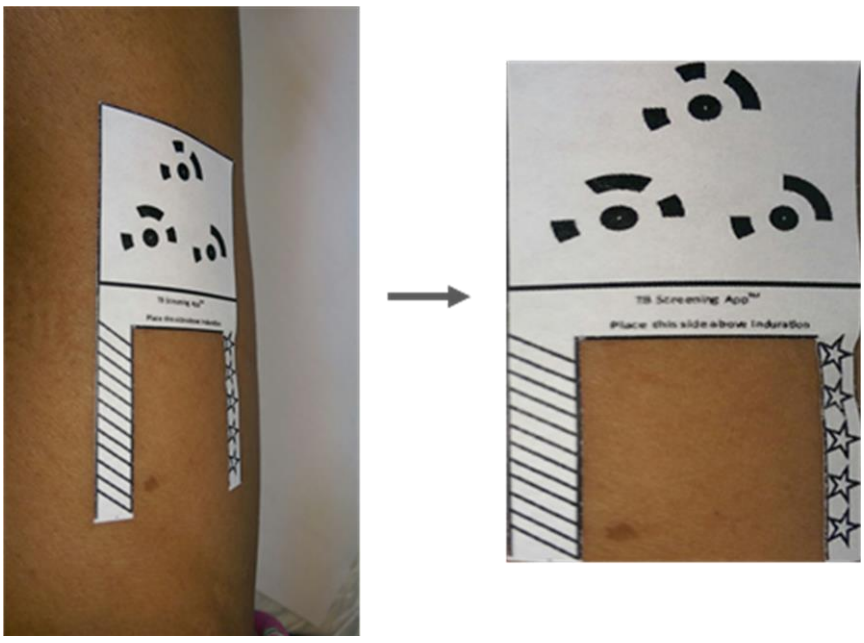


Figure 5.6: Result from isolating and transforming the scaling sticker in an induration image.

Similarly, to how the scaling sticker was found and isolated in the original image, the induration box was found and extracted from the scaling sticker. It was resized to 1000x1200 pixels to normalise the size and resultant diameter measurements. The result is shown in Figure 5.7.

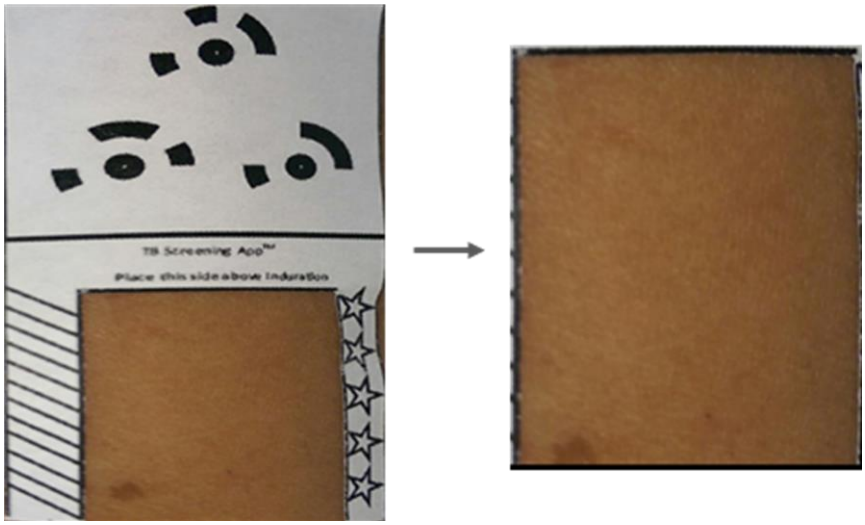


Figure 5.7: Results from extracting the induration box from the scaling sticker.

After isolation and normalisation of the induration boxes, the domes were delineated to allow the diameter measurement to be extracted. The following steps were taken in extracting the diameter measurement of a dome from the normalised induration box image:

a) Segmentation

   To segment the dome, a random walker segmentation method was used (Grady, 2006). The seed areas for the algorithm (bounding the area where the segmentation will occur) were specified to be the same in all the images, and to be to be two concentric circles with their centres being at the centremost pixel of the image. The region between the circles was then segmented to delineate the induration.

b) Contour Finding and Ellipse Fitting

   The OpenCV contour finding algorithm (cv2.findContours) was applied to find the segmented area in the image, the contour corresponding to the dome. The ellipse fitting algorithm (cv2.fitEllipse) was then used to plot an ellipse of best fit to the contour with reference to all

the points in the contour. This gives a uniform elliptical shape from which the diameter is measured.

c) Diameter calculation

Because the induration images boxes were all normalized to a size of 1000x1200 pixels, the true diameter of the ellipse could be calculated. The induration box on the scaling sticker is 20mm wide. Therefore, by extracting the ellipse diameter in pixels from the previous algorithm, and dividing it by 50, the true diameter measurement of the dome could be determined in millimetres.

5.3 Experiments

Initially, to test the effectiveness of the induration image box extraction algorithms, they were run on all 63 of the available real induration images, excluding the images taken at an angle of 40° or wider from the axis perpendicular to the arm as at this angle the induration was not visible at all. Of the 63 assessed images, 57 of the induration boxes were correctly identified, cropped and transformed. This translates to an induration box identification and transformation accuracy of 90%. The errors in the 6 images which were not segmented correctly arose from the guidance components on the application interface not being adhered to. This, for example, caused the scaling sticker to appear larger than expected in the original image causing the algorithms to fail.

For testing the effect of the image angle and orientation during image assessment on the accuracy of the assessment, a distinct synthetic alternative was used in place of the real indurations to eliminate segmentation errors which could arise as a result of the real indurations being unclear and difficult to identify. This provided a robust way to isolate the effects of the changes in angle and orientation of the images, and test whether it affects the diameter measurement of the object being assessed, without being affected by inaccurate segmentation of the object.

For the initial assessment, the diameter of a white sticker was assessed as it was in good contrast to the skin of the subject. The results from measuring the diameter of the white sticker, in images taken at varying orientations and angles, are shown in Figure 5.8.

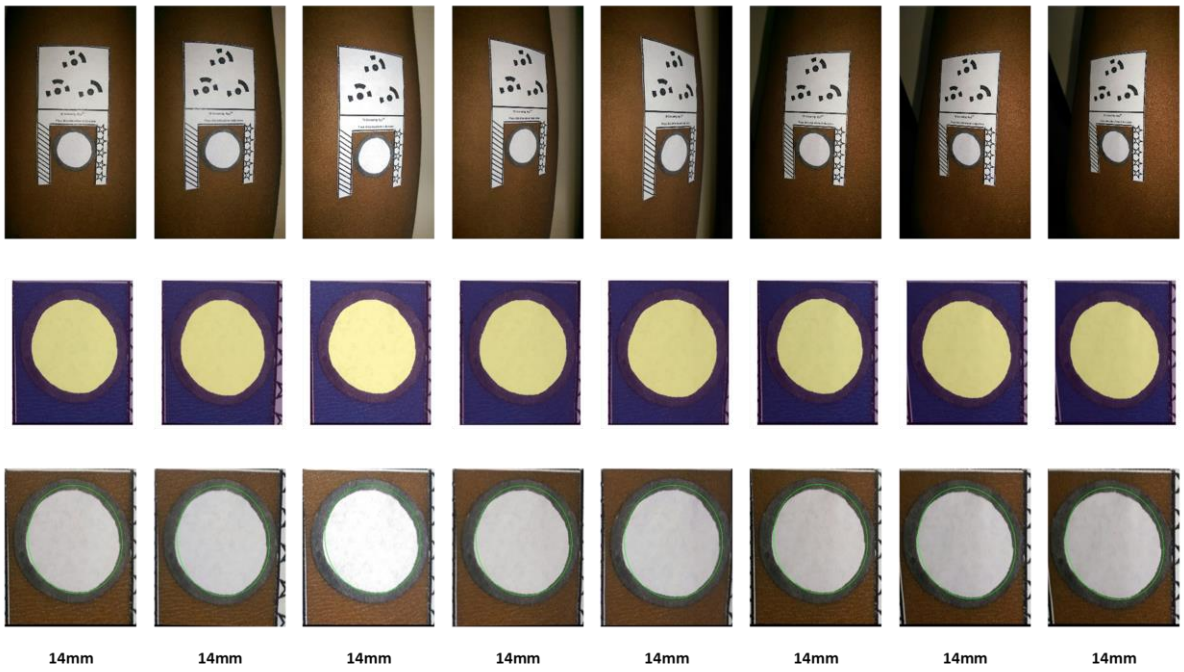


Figure 5.8: Sticker delineation and diameter measurement results for images of stickers taken at varying orientations above the subject's arm.

Despite the images being taken at a variety of angles over a range of 70°, every resultant diameter measurement was 14mm. It is therefore correct to conclude that the orientation of an image does not affect the diameter measurement of a flat object if the image has undergone a perspective transformation for size and shape normalisation.

To test whether the results would be consistent with those obtained during the assessment of a raised object, artificial white domes of varying height were used as the object being assessed. A maximum height of 4mm was used as a height of more than 4mm is unlikely for an induration. The domes actual diameter measurements are shown in Table 5.1.

Table 5.1: Actual diameter measurements for each of the raised domes used to test if image orientation affects the diameter measurement taken from images of an object.

Height	Diameter
1 mm	13 mm
2 mm	12 mm
3 mm	12 mm
4 mm	13 mm

These domes are shown in Figure 5.9. Similarly, to how the sticker was segmented and measured, the domes were assessed. The resultant diameter measurements for each of the domes and their respective image orientations, can be seen in Table 5.2.

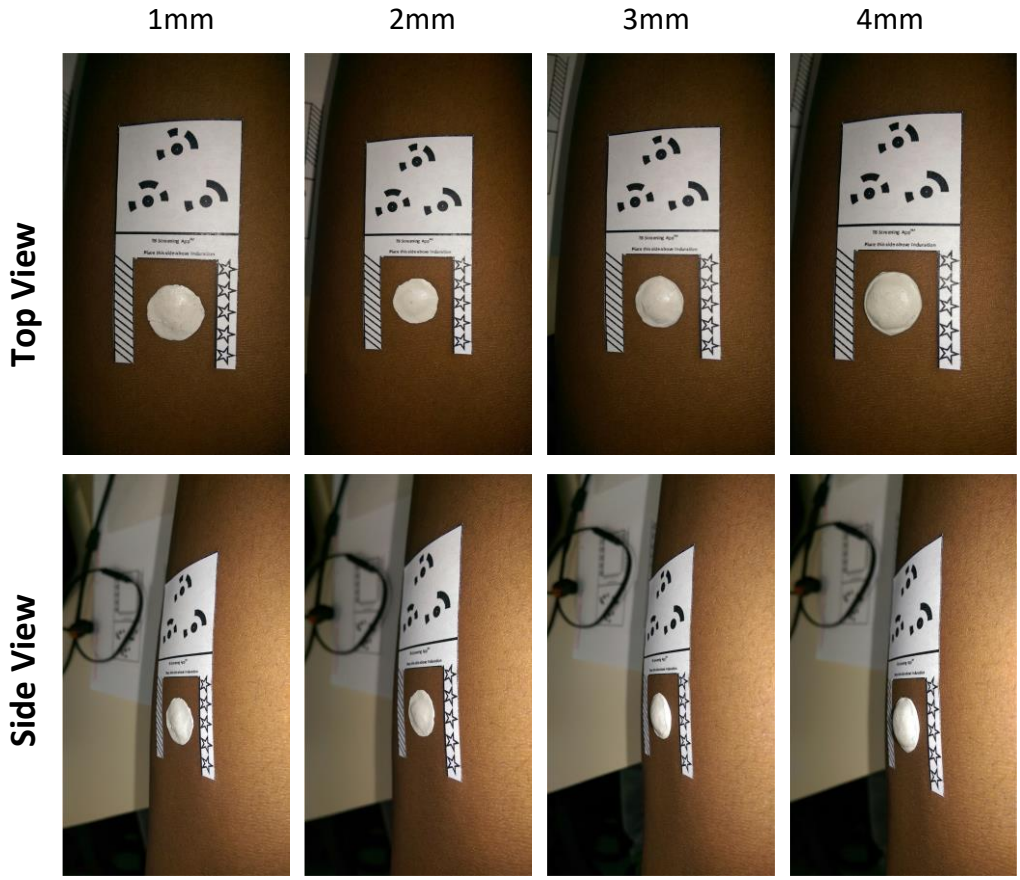


Figure 5.9: Top and side view images of domes with heights of 1 mm,2 mm,3 mm and 4 mm used to test whether the height of the object being assessed affects the diameter measurement.

Table 5.2: Diameter measurement results in millimetres from images of domes of varying height taken at varying orientations above the subject's arm.

Dome Height	Image orientation (Angle from plane perpendicular to the long axis of the arm)							
	-40°	-30°	-20°	-10°	0°	10°	20°	30°
1mm	13	13	13	13	13	13	14	14
2mm	12	12	12	12	12	12	12	13
3mm	12	12	12	12	12	12	13	13
4mm	13	13	13	13	13	13	14	14

The resultant diameter measurements from the assessment of the domes did not vary by more than one millimetre for each of the domes. This indicates that even if the object being assessed is raised

up to 4 mm, the variation in height does not substantially affect the 2D diameter measurement results.

5.4 Discussion

In this chapter, the need for the 3D reconstruction aspect of the LTBI screening tool was assessed. Experiments carried out showed that the induration could be analysed more effectively using a single 2D image of an induration. This is because less information is lost in this method compared to the 3D reconstruction procedure. Using a single 2D image for analysis is also cost-effective since only one image, instead of several images, would be transmitted from the mobile device to the remote processing centre and the reduction in the number of images needed to be captured increases the usability of the application. Furthermore, the experimental results obtained also showed that the orientation of the phone camera during image capturing does not affect the diameter measurement of a flat object, if the image has undergone a perspective transformation for size and shape normalisation. Additionally, if the object being assessed has a height of up to 4 mm, it does not affect the 2D diameter measurement results substantially.

6 Synthetic induration image generation using a generative adversarial network

As an alternative image assessment method to the 3D reconstruction and 2D image processing methods used in the first version of the LTBI screening tool, induration image classification using a neural network is worth considering. It is hypothesized that deep learning-based image classification may detect salient features in the induration images that could be used to classify the induration into the categories analogous to the induration diameter categories of the WHO as shown in Table 2.1. However as mentioned in the literature review, one of the main limitations of deep learning, which has significance for many medical applications (including this one), is the high dependency on large amounts of data, in this case, large numbers of images. This is due to medical images being costly to acquire and the large amount of work needed to be done by experts to produce and label the medical images. Furthermore, the need to ensure the confidentiality of patient information and the integrity of the medical data also limits the number of medical images that can be collected for LTBI screening using the proposed screening tool.

This chapter therefore describes an effort towards circumventing the collection of large real datasets to be used in the LTBI screening tool. This was achieved by developing artificial datasets to enable an alternative deep learning-based image assessment method. The generation of artificial datasets is based on the use of a generative adversarial network (GAN), as introduced by Goodfellow et al. (2014), for synthetic image generation. Generative adversarial networks make it possible to produce synthetic images to increase the size of the induration image dataset, and therefore reduce the number of real induration images needed to build a deep learning-based image classification network for the LTBI screening tool.

6.1 Technical considerations of a generative adversarial network

A GAN consists of two neural networks, one acting as a generator, which generates synthetic images, and the other acting as a discriminator, which tries to identify the generated images as synthetic compared to the training images (Goodfellow et al., 2014). The generator and discriminator contest with each other until a sample designed by the generator is indistinguishable for the discriminator from the true samples.

As shown in Figure 6.1, the generator takes noise z as an input, which is passed into the network from the latent space (a vector with specific dimensions) and is used by the generator G to generate the synthetic images x ($x = G(z)$). The synthetic images are compared by the discriminator against the real training images where the discriminator aims to classify the images as real or synthetic. The output of the discriminator $(D(x))$ is a probability, between 0 and 1, of the image x with 1 representing an authentic image and 0 representing a synthetic image. The goal of the discriminator is therefore to maximize the chance of recognizing real images as real and the synthetic images as fake. The discriminator's loss in achieving this goal is measured using two logarithmic functions as shown in Equation 1, one for the real images and reversing the label for the synthetic images, to allow for a function with a max value objective (Goodfellow et al., 2014). The logarithmic function using the discriminator's probability result is multiplied by the expected value over all real data instances ($\mathbb{E}_{x\sim p_{data}(x)}$) and the logarithmic function using the generator's probability result is multiplied by the expected value over all the generated synthetic instances ($\mathbb{E}_{z\sim p_z(z)}$).

$$\max_D \; V(D) = \; \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \; \mathbb{E}_{z\sim p_z(z)}[\log(1 - D\big(G(z)\big))] \qquad \text{[Eq. 1]}$$

The objective function on the generator side aims to generate images with the highest possible value of $D(x)$ as this would mean it has been successful in fooling the discriminator into classifying the synthetic images as real. Considering the generator influence in the function above it tries to achieve the highest value of $D(x)$ by minimizing the function in Equation 2 (Goodfellow et al., 2014).

$$\min_G V(G) = \; \mathbb{E}_{z\sim p_z(z)}[\log(1 - D\big(G(z)\big))] \qquad \text{[Eq. 2]}$$

The contradictory objective functions of the generator and the discriminator, where G tries to minimize V and D tries to maximize it, and the overall aim of the GAN are therefore described by the formula in Equation 3 (Goodfellow et al., 2014).

$$\min_{G} \max_{D} V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad \text{[Eq. 3]}$$

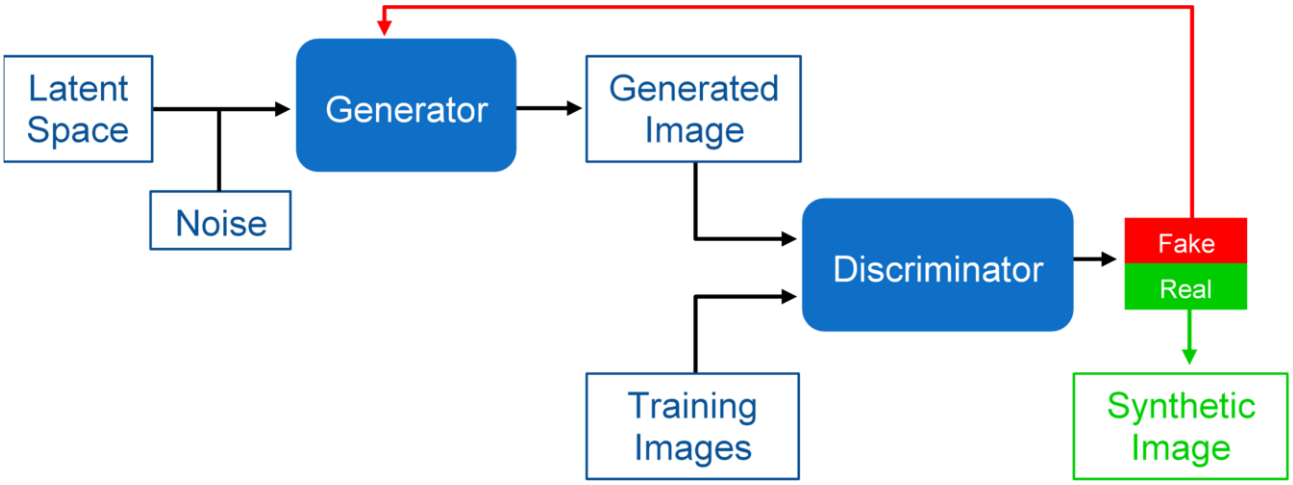


Figure 6.1: Generative adversarial network structure illustrating the generator and discriminator models and how they interact to produce realistic synthetic images (Goodfellow et al., 2014).

6.2 Structure of a robust GAN for synthetic induration image generation

When designing the GAN to generate synthetic induration images, five main modules were considered: data pre-processing and loading, the generator, the discriminator, the loss function and the training module.

a) Data pre-processing and loading

The training images needed to be transformed into the correct format and size before they could be passed into the discriminator network. This ensured that the format of the real and the generated images is the same for a fair comparison by the discriminator. To remodel the images to be compatible to the network, they were transformed in all of the following ways:

- Resizing the image to the required size and number of pixels.
- Converting the pixels in the image array from Red Green Blue (RGB) values between 0 and 255 to a tensor with values between 0 and 1. A tensor here, is a 3-dimensional matrix which generalises the 3 colour values associated with each pixel in an image. This was done to enable faster convergence during training of the model.

- Normalising the tensor values to be between -1 and 1. This was necessary for the utility of activation function layers in the network as they require both negative and positive inputs to produce required results.

Subsequently to the images being transformed, they were passed into a data loader. The data loader was used to retrieve and load the images of the training set in batches of a specific size into the network. It was also used to shuffle the images before they are split into batches to ensure that there is a wide variety of images in each batch.

b) Discriminator

The role of the discriminator as mentioned in section 6.1 was to distinguish between the synthetically generated images and the real images. Hence, the architecture of a discriminator was similar to a traditional convolutional neural network (CNN) as it performed a simple classification task between the two classes of images. Similarly to how a CNN may be trained to classify images as either a dog or a cat, the discriminator was trained to classify images as real or synthetic. The input of the discriminator was therefore either synthetic images generated by the generator or real images from the training dataset.

The architecture of the discriminator comprised four types of layers as described in section 2.3 of the literature review; convolution, pooling, normalisation and activation. The induration images had two spatial dimensions, height and width; therefore, the network layers, through which the images are passed, also needed to be in the 2D form to allow for convolution, normalisation and pooling to occur along both axes (height and width).

The four layers present in the discriminator are discussed below:

- 2D convolution applied sliding convolutional filters to the input image.
- Activation function broke linearity in the network. This was necessary to allow the network to learn complex non-linear functional mappings from the data.
- 2D batch normalisation normalized all the features in the input channel of the layer along the dimension of the batch.
- 2D pooling performed down-sampling by dividing the input tensor into rectangular pooling regions and computing the maximum or average values of each region.

c) Generator

Similarly, the architecture for the generator also contained a sequence of layers. Because the discriminator was a CNN, the generator was designed to be a deconvolutional neural network. Deconvolution is also known as the transpose of the convolution and involves up-sampling of data (Zeiler & Fergus, 2014). Contrary to normal convolution, which forms a many-to-one relationship, where images are compressed into feature layers, transposed convolutions work in the opposite direction having a one-to-many relationship where each pixel is expanded into a matrix with each value in the new matrix having a specific weighting. With image generation, which was the role of the generator, there was a need for up-sampling the noise vector input into high resolution images.

The four layers present in the generator are discussed below:

- 2D transposed convolution up-sampled the feature maps with sliding transposed convolutional filters.
- Activation function - same function as in discriminator module.
- 2D batch normalisation - same function as in discriminator module.
- 2D pooling - same function as in discriminator module.

d) Loss function and optimisation

To determine the error in the prediction given by the discriminator and subsequently optimise the training of the network considering this error, a loss function, as shown in Equation 3, was used. Optimisers were also created for both the discriminator and the generator to update the network weights during training, based on the loss from the loss function. The Adam optimiser was used as it yields the most realistic results for image generation tasks and does so in an efficient and fast manner (Kingma & Ba, 2017).

e) Training

The training module of the neural network iterated over all the images in the dataset, a set number of times (epochs), and in set batch sizes. During this process the weights in the layers of the neural network were updated and eventually set to the optimum values which produced realistic-looking images.

The training process involved all of the following steps:

- Training of the discriminator using the real image dataset as well as the synthetic images generated by the generator.
- Using the loss function to determine the error in prediction given by the discriminator.
- Backpropagating this error with respect to the weights in each layer of the discriminator, using the optimiser to update the weights according to how much each weight is responsible for the overall error.
- Similarly to how the discriminator weights were updated, the generator weights were also updated considering the overall error found for the generator.
- These steps were iterated through as many times as the 12-hour time constraint given by Google Colaboratory allowed, to find the most suitable weights to produce synthetic images that resemble real induration images when compared using the Fréchet inception distance (FID) evaluation metric. The loss values were printed at the end of each iteration to monitor the success of the training throughout the training process.
- The final generated images were then saved after completion of the network training.

6.3 Tuning GAN hyperparameters to accommodate limited image datasets

For a GAN to produce realistic looking images, hyperparameters need to be tuned to best suit each application. These hyperparameters should be chosen based on the size of the training image dataset. The following hyperparameters in the GAN architecture were therefore tailored to allow for better training when using the small induration dataset. Section 6.7 details the values used while tuning these parameters.

a) The Batch Size

Larger batch sizes usually result in faster training of the network. Using smaller batch sizes results in slower training but often results in faster convergence of the network specifically if training is done on a small dataset.

b) Epochs

   Generally, with each epoch a model becomes more accurate, up to a point where is starts to plateau (converge). Therefore, when using a limited dataset, it will usually take longer to converge, thus a high number of epochs should be used.

c) Learning rate

   A low learning rate is more reliable as small steps are taken towards the minimum of the loss function and there is less chance of the minimum being overshot, but as a result, optimisation takes longer. With a small training dataset, a low learning rate should be used to ensure that no minima are overlooked despite the fact that the optimisation will take longer.

Overall there is a trade-off between the size of the dataset and the time taken to train a neural network. If the training dataset is very limited, the batch size, number of epochs and learning rate should all be applied as optimally as possible to allow for as much information to be extracted from the training data as possible, rather than trying to optimise the time taken to train the network. These points were all considered while tuning the hyperparameters of the GAN detailed in Section 6.7.

6.4 Final GAN framework for generating synthetic induration images

Google Colaboratory was chosen as the platform on which to build and train the GAN network as it satisfied the computing requirements for the task.

An existing GAN architecture was chosen as a starting point for the final GAN framework on which the key components of a robust GAN framework discussed in Section 6.2 were applied. The chosen architecture guidelines came from the deep convolutional generative adversarial network (DCGAN), as this architecture had shown good performance on unsupervised tasks, specifically with colour image generation tasks (Radford, Metz & Chintala, 2016). The DCGAN's structure can be seen in Figure 6.2.

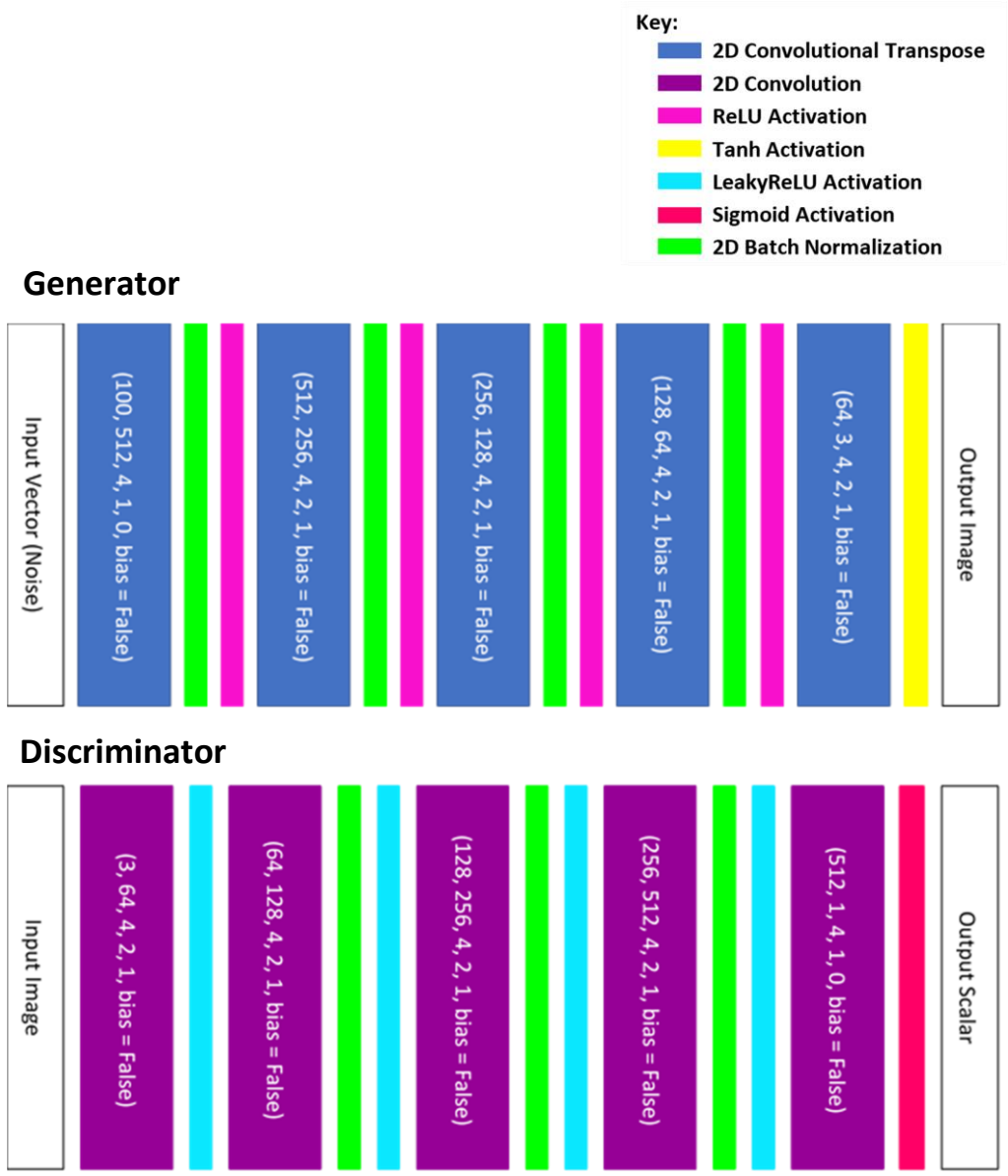


Figure 6.2: Layers in the generator and discriminator modules of the DCGAN framework.

Both the generator and discriminator have six parameters set in each of their respective convolution layers. The first parameter is the number of input channels, this varies depending on what the input to the layer is and is determined from output of the preceding layer. The second parameter is the number of channels produced as an output from the convolution in that layer. Both of these values therefore change for each layer. The last layer in the generator module has an output of 64 channels which results in the generated images being designed as a standard size of 64x64 pixels. The subsequent four parameters of the convolution layers remain the same; the convolution kernel is set to 4x4 pixels, the stride of the convolution is set to 2 pixels, padding between the resultant pixels

from each stride is set to 1 pixel and lastly the bias is set to false to prevent a learnable bias being added to the output of each convolution layer. However, there are two exceptions to these consistent parameters; the padding parameter is set to 0 rather than 1 in the first layer of the generator as well as in the last layer of the discriminator to prevent checkerboard artefacts (uniform horizontal and vertical lines) appearing in the original input image and the final output image. Checkerboard artefacts are a well-known consequence of the deconvolution process in the generator module of the GAN (Sugawara, Shiota & Kiya, 2018).

The DCGAN framework was populated with the GAN components discussed in Section 6.2. Experiments were then done to tune the hyperparameters and find values which produced the best results when training was done using a limited dataset. These experiments are detailed in the following section.

6.5 GAN implementation

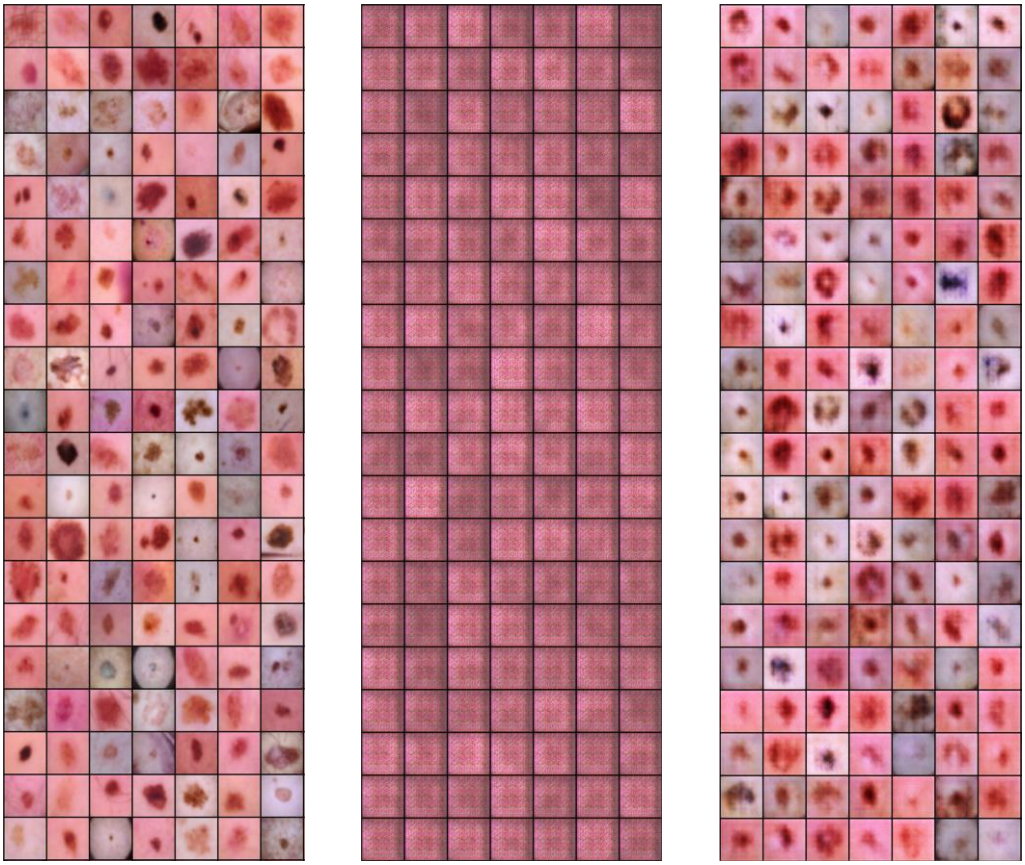
Initially, to avoid limitations due to insufficient data, the GAN was trained to produce synthetic images using the HAM10000 dataset (Tschandl, Rosendahl & Kittler, 2018). The HAM10000 dataset, is a large collection of multi-source dermatoscopic images of common pigmented skin lesions, which contains 10000 training images and 11788 testing images. The dataset was chosen to train the initial model since the skin pigmentation images have many similar features to the induration images. Both image datasets represent close range 2D images of features on human skin. The 'noise', such as hairs and freckles, are the same in both sets of images. Furthermore, the composition of the images is very similar, both have a large central feature (induration or melanoma) surrounded by human skin with no other objects in the background to complicate the feature identification during image generation.

6.6 GAN evaluation metric

The FID score was chosen as the evaluation metric to measure the performance of the GAN image generation. The goal of the GAN is to produce synthetic images which closely resemble real induration images. The FID evaluation metric was therefore chosen for the this application as it was most suitable in that it compares the synthetic image dataset to a target image dataset rather than evaluating the synthetic image dataset in isolation as discussed in section 2.3.5 (Heusel et al., 2017).

6.7 Results

The GAN was initially trained using all 10000 images, with a batch size of 200 and using 10 epochs. The number of epochs was then increased to 30. These values were chosen by trial and error as there is currently no gold standard for the number of epochs to be used to train a GAN. As can be seen in Figure 6.3, by increasing the number of epochs to 30, the generated images appear much more realistic compared to when 10 epochs were used.



Real images **Synthetic images (10 epochs)** **Synthetic images (30 epochs)**

Figure 6.3: Real melanoma images compared to synthetic melanoma images produced using a batch size of 200 with 10 and 30 epochs.

The number of images used from the HAM10000 was then reduced to 200 images to imitate the limited induration image dataset. The batch size was decreased to 25 to accommodate for the small dataset. The number of epochs was increased to 100, 200 and 400. These epoch values were again chosen by trial and error due to the lack of a current gold standard value. The trade-off for more epochs is an increase in processing time. Therefore, in this work, a maximum number of 400 epochs was chosen to allow processing to completed within the 12-hour limit per training session given by Google Colaboratory. The learning rate was set at a constant value of $2x10^{-4}$.

The graph in Figure 6.4 illustrates how the generator loss decreased rapidly after approximately 75 epochs as the model began to produce realistic looking induration images. This illustrates when convergence took place.

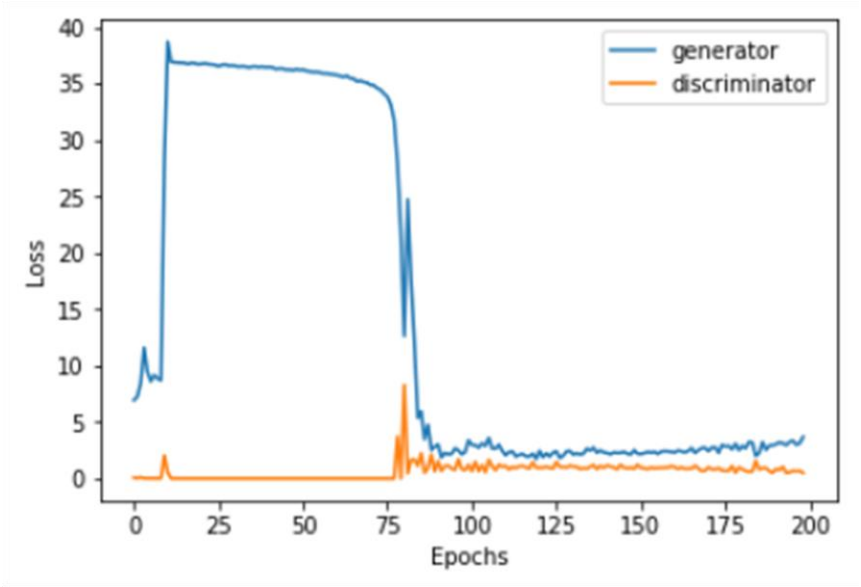
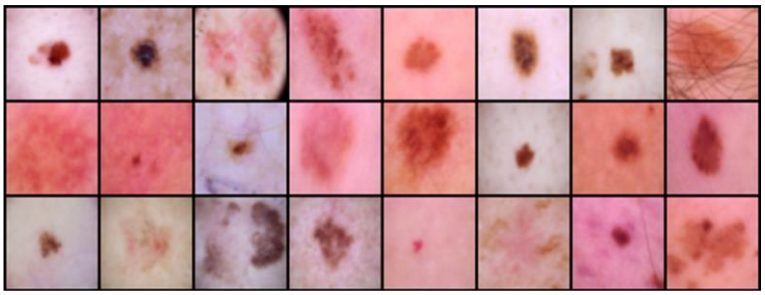
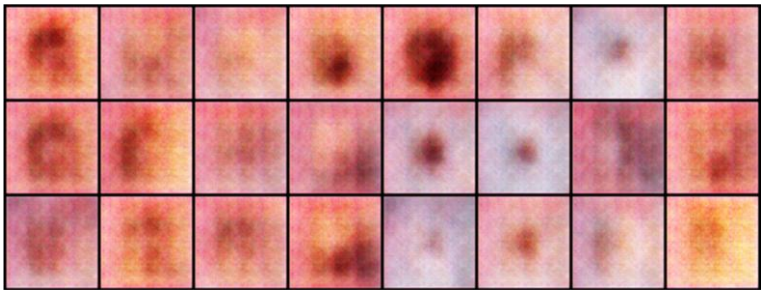


Figure 6.4: Convergence of the generator (blue) and discriminator (orange) network losses as the number of training epochs increases.

The resultant synthetic images produced by the GAN after 100, 200 and 400 epochs are shown in Figure 6.5.

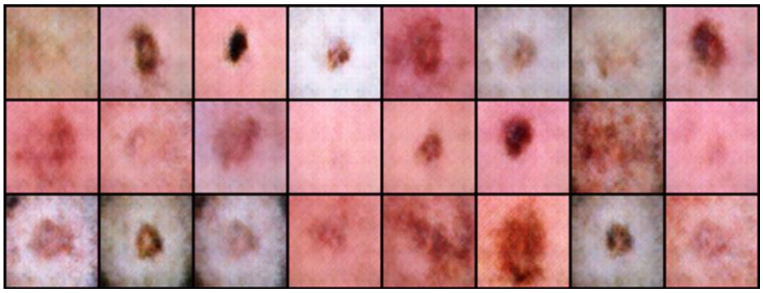


Real melanoma images



Synthetic images (100 epochs)

Synthetic images (200 epochs)



Synthetic images (400 epochs)

Figure 6.5: Real melanoma images compared to synthetic melanoma images produced using a limited dataset of 200 images, batch size of 25, learning rate of $2x10^{-4}$ with 100, 200 and 400 epochs.

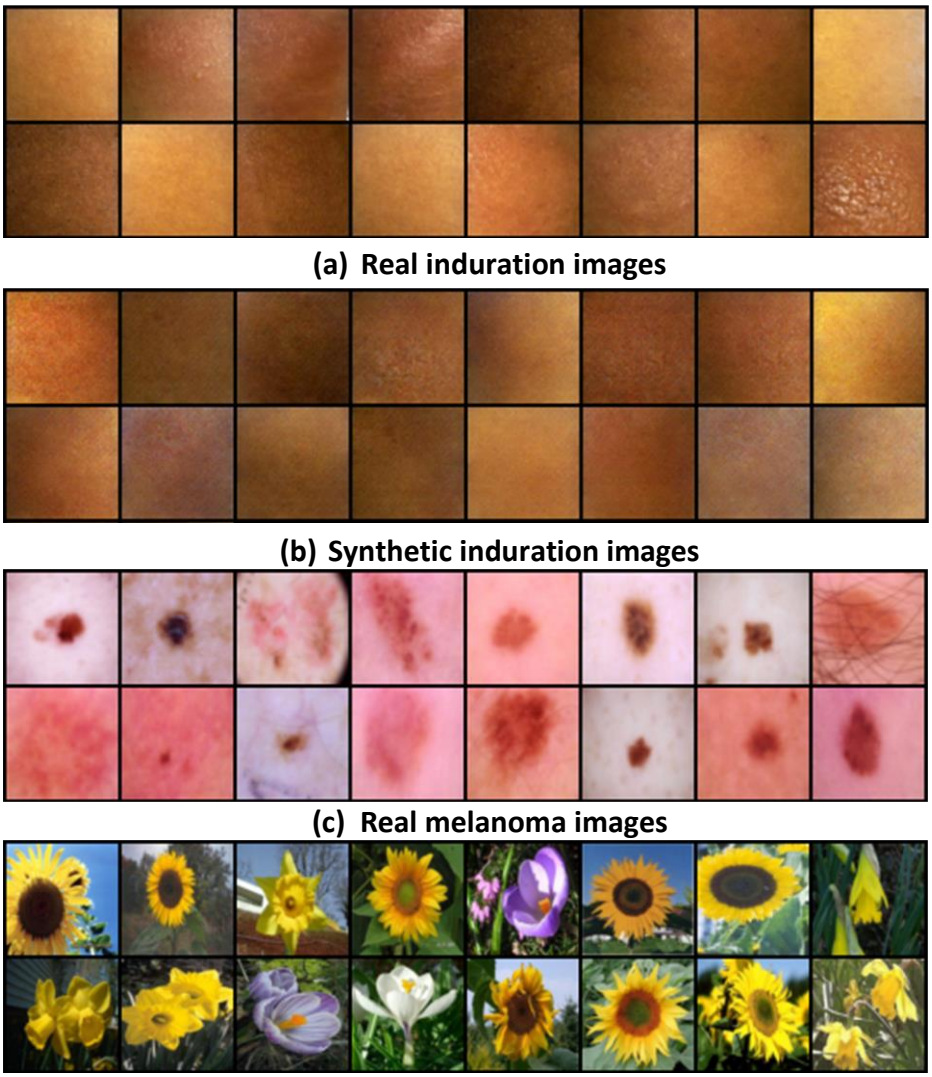
The FID Scores for each of the groups of images in Figure 6.5, when compared to a dataset of real melanoma images, are shown in the table below:

Table 6.1: FID Scores for melanoma image groups when compared to a dataset of real melanoma images.

Image Group	**FID Score**
Synthetic melanoma images (100 epochs)	263.09
Synthetic melanoma images (200 epochs)	227.10
Synthetic melanoma images (400 epochs)	213.89
Real melanoma images	*111.59*

The results in Table 6.1 show that as the number of epochs increases, the corresponding FID scores decrease, confirming the visual assessment that the synthetic images became more realistic with an increasing number of epochs.

Considering that the best results were obtained for the melanoma images when using 400 Epochs. The GAN framework was tested and trained for 400 Epochs using the 150 available induration images. The synthetic induration images generated by the designed GAN framework appear to be very similar to the real induration images as can be seen in Figure 6.6. Figure 6.6 also shows real melanoma images and images of real flowers. The FID score for each of the groups of images in Figure 6.6 when compared to real induration images is shown in Table 6.2. As can be seen from the results, the synthetic induration images (with a FID score of 90.73) are more similar to the real induration images (with a FID score of 36.66) relative to the other two groups of images (with FID scores of 332.39 and 480.04).



(a) Real induration images

(b) Synthetic induration images

(c) Real melanoma images

(d) Real flower images

Figure 6.6: (a) Real induration images, (b) synthetic induration images generated by the designed GAN framework using 150 induration images for training, (c) real melanoma images and (d) real images of flowers.

Table 6.2: FID Scores for real induration images, synthetic induration images, melanoma images and images of flowers when compared to a dataset of real induration images.

Image Group	**FID Score**
Real induration images	*36.66*
Synthetic induration images	90.73
Melanoma images	332.39
Flower images	480.04

6.8 Discussion

In this chapter, synthetic induration images were generated using a GAN from a limited induration image dataset. The adapted DCGAN architecture proved to be a suitable framework for this work as it produces good results during colour image generation. Furthermore, the DCGAN contained hyperparameters which were manipulated to allow for successful synthetic image generation to be achieved using a very limited training dataset. These hyperparameters were; the batch size, the number of epochs and the learning rate. Changing these hyperparameters enabled realistic looking synthetic images to be produced from a dataset as small as 150 images, with the most effective change to the hyperparameters being increasing the number of epochs.

Figure 6.5 and Figure 6.6 show synthetically generated melanoma and induration images respectively, produced by the GAN framework. The success of the GAN in generating these synthetic images was analysed using the Fréchet inception distance score. The results highlight the potential for synthetic induration image generation using limited datasets which could enable a deep learning-based induration image classification model to be trained.

A limitation of the generative adversarial network used in this work, is that the generated synthetic images are low resolution images. The network architecture only allows for images of size 64x64 pixels to be generated, and smaller salient details in the images may be lost. Research into a super resolution network which either generates images of higher resolution or increases the resolution of the small images should be considered.

6.8.1 Considerations for a deep learning based LTBI screening tool

The potential of a deep learning based LTBI classification model should be assessed once a large dataset has been generated to train the model. Furthermore, it is suggested that the accuracy and comprehensiveness of the synthetic images in containing all the salient characteristics present in real induration images should be assessed before they are used to train a classification model.

Classification of induration images based on a deep learning model, would require measured induration diameters of real induration images to be incorporated into the training of the system. The ultimate test of the fidelity of synthetic images, would be the in the ability of a the classification

system to correctly place test images of real indurations in categories equivalent to the diameter categories of Table 2.1.

Zeiler & Fergus (2014) introduce a method for visualizing the features used by the network to guide a classification which could give insight into the salient features of induration images used during classification. This knowledge could be used to improve image acquisition conditions to ensure that the salient features are captured. The insights provided by the deep learning model might aid in finding features of the induration, other than the diameter, to use in classification.

As an alternative approach to generating synthetic induration images, to overcome the limitation of a small image dataset, a method introduced by Fei-Fei, Fergus & Perona (2006), known as one-shot learning, could be explored. The intention of one-shot learning is to learn information about image categories from one, or very few training images as opposed to the machine learning categorization approach which requires a large image dataset for training. This categorization approach is demonstrated by humans in how they use prior knowledge about objects to assist in classifying new objects without having seen numerous different examples of the object.

7 Conclusion

The aim of this study was to evaluate the backend image analysis framework of the UCT mHealth app and to explore an alternative approach to induration assessment through deep learning. It has become evident that the real indurations in the images captured using the UCT mHealth app are much flatter and less visible compared to the mock indurations used to evaluate the earlier version of the LTBI screening tool. As a result, the algorithms in the earlier screening tool were unsuccessful in identifying and delineating real indurations. This is because the 3D reconstruction procedure was unable to highlight the induration, which nullifies the succeeding assessment steps as they use the resultant 2D depth map image produced by the 3D reconstruction procedure. The purpose of the 3D reconstruction, besides highlighting the induration, was also to ensure that the induration was assessed from a normalised image orientation. From the analysis in Chapter 5, it is evident that the position at which images are taken of an induration does not affect the resultant diameter measurement of the induration, if the image undergoes a perspective transformation before the assessment. We could therefore conclude that the 3D reconstruction aspect of the assessment is unsuccessful in highlighting the induration for segmentation and is trivial in normalising the orientation and angle from which the induration is assessed.

Due to the lack of success in induration delineation using 3D reconstruction methods and standard 2D image processing, an alternative assessment for induration image assessment, using deep learning, was considered. A limitation to this approach is the lack of a large induration image dataset to train the classification network. A generative adversarial network (GAN) was therefore proposed to generate synthetic induration images in an attempt to increase the size of in induration image dataset. The success of the network was tested using the Fréchet inception distance score which suggested that the generated induration images resembled real induration images despite being trained on a dataset of only 150 induration images. This highlights the potential of synthetic induration image generation, using a limited dataset, to produce enough images to train a deep learning image classification model to classify induration images. Such classification, combined with other patient information, could aid in detecting the presence of LTBI.

Although the GAN was able to produce synthetic images which closely resemble real induration images, the images have not been assessed to determine if they contain all the salient characteristics

of real induration images. It is suggested that the accuracy and comprehensiveness of the synthetic images be assessed in future work. Subsequently, a classification network should be trained using the synthetic images and tested on a dataset of real induration images to determine the viability of using synthetically generated images for a deep learning classification approach for the LTBI screening tool analysis.

The low resolution of the synthetic induration images is a limitation, and alternative GAN tools should be considered, which can accommodate higher resolution images. One-shot learning could be considered as an alternative approach to generating synthetic induration images, to overcome the limitation of a small image dataset for a deep learning classification model.

8 References

Al-Orainey, I.O. 2009. Diagnosis of latent tuberculosis: Can we do better? *Annals of Thoracic Medicine.* 4(1):5-9. DOI:10.4103/1817-1737.44778.

Albabtain, A.F., AlMulhim, D.A., Yunus, F. & Househ, M.S. 2014. The Role of Mobile Health in the Developing World: A Review of Current Knowledge and Future Trends. *Journal of Selected Areas in Health Informatics.* 4(2):10-15.

Baur, C., Albarqouni, S. & Navab, N. 2018. Generating Highly Realistic Images of Skin Lesions with GANs. DOI:10.1007/978-3-030-01201-4_28.

Borji, A. 2019. Pros and cons of GAN evaluation measures. *Computer Vision and Image Understanding.* 179:41-65. DOI:10.1016/j.cviu.2018.10.009.

CDC. 2013. Mantoux Tuberculin Skin Test Chart. Available: https://www.cdc.gov/tb/publications/posters/images/Mantoux_wallchart.pdf [8/31/2018]. Available: https://www.cdc.gov/tb/publications/posters/images/Mantoux_wallchart.pdf.

Dendere, R., Mutsvangwa, T., Goliath, R., Rangaka, M.X., Abubakar, I. & Douglas, T.S. 2017. Measurement of Skin Induration Size Using Smartphone Images and Photogrammetric Reconstruction: Pilot Study. *JMIR Biomedical Engineering 2.* 1(e3). DOI:http://dx.doi.org/10.2196/biomedeng.8333.

Douglas, D. & Peucker, T. 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization.* 10(2):112-122. DOI:10.3138/FM57-6770-U75U-7727

Ellner, J. 2016. Tuberculosis. In *Goldman-Cecil Medicine (25th Ed.).* Philadelphia: Elsevier/Saunders. 2030-2039. DOI:10.1016/B978-1-4377-1604-7.00332-8.

Fei-Fei, L., Fergus, R. & Perona, P. 2006. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 28(4):594-611. DOI:10.1109/TPAMI.2006.79.

Glaziou, P., Sismanidis, C., Floyd, K. & Raviglione, M. 2015. Global Epidemiology of Tuberculosis. *Cold Springs Harbor Perspectives in Medicine.* 5(2). DOI:10.1101/cshperspect.a017798.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems.* 2:2672- 2680. arXiv preprint arXiv:1406.2661v1.

Grady, L. 2006. Random Walks for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 28(11):1768-1783. DOI: 10.1109/TPAMI.2006.233

GSMA, I. 2017. *Global Mobile Trends 2017.*

Haghdoost, A.A., Afshari, M., Baneshi, M.R., Gouya, M.M., Nasehi, M. & Movahednia, M. 2014. Estimating the Annual Risk of Tuberculosis Infection and Disease in Southeast of Iran Using the Bayesian Mixture Method. *Iranian Red Crescent Medical Journal.* 16(9). DOI:10.5812/ircmj.15308.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*. arXiv preprint arXiv:1706.08500.

Huebner, R.E., Schein, M.F. & Bass, J.B. 1993. The tuberculin skin test. *Clinical infectious diseases.* 17(6):968-975. DOI: 10.1093/clinids/17.6.968

ICASA. 2018. *Q2 Bi Annual Retail Tariffs Report Jan Jun 2018*. Sandton.

Karageorgos, G., Andreadis, I., Psychas, K., Mourkousis, G., Kiourti, A., Lazzi, G. & Nikita, K.S. 2018. The Promise of Mobile Technologies for the Health Care System in the Developing World: A Systematic Review. *IEEE reviews in biomedical engineering.* DOI:10.1109/RBME.2018.2868896.

Karras, T., Aila, T., Laine, S., Lehtinen, J. & Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv preprint arXiv:1710.10196.

Kim, Y.-T. 1997. Contrast enhancement using brightness preserving bi-histogram equalization. *IEEE Transactions on Consumer Electronics.* 43(1):1-8. DOI:10.1109/30.580378.

Kingma, D. & Ba, J. 2017. Adam: A Method for Stochastic Optimization. *3rd International Conference for Learning Representations*. San Diego, 2015. arXiv preprint arXiv:1412.6980.

Lakhani, P., Gray, D., Pett, C., Nagy, P. & Shih, G. 2018. Hello World Deep Learning in Medical Imaging. *Journal of Digital Imaging.* 31(3):283-289. DOI:10.1007/s10278-018-0079-6.

Latif, S., Rana, R., Qadir, J., Ali, A., Imran, M.A. & Younis, M.S. 2017. Mobile health in the developing world: Review of literature and lessons from a Case study. *IEEE Access.* 5:11540-11556. DOI:10.1109/ACCESS.2017.2710800.

Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G.B., Seo, J.B. & Kim, N. 2017. Deep Learning in Medical Imaging: General Overview. *Korean journal of radiology.* 18(4):570-584. DOI:10.3348/kjr.2017.18.4.570.

Linas, B.P., Wong, A.Y., Freedberg, K.A. & Horsburgh, C.R. 2011. Priorities for screening and treatment of latent tuberculosis infection in the United States. *American journal of respiratory and critical care medicine.* 184(5):590-601. DOI:10.1164/rccm.201101-0181OC.

Manosuthi, W., Wiboonchutikul, S. & Sungkanuparph, S. 2016. Integrated therapy for HIV and tuberculosis. *AIDS Research and Therapy.* DOI:10.1186/s12981-016-0106-y.

Mardani, M. & Abtahian, Z. 2015. New Advances in Diagnosis of Latent Tuberculosis Infection: A Review Article. *Archives of Pediatric Infectious Diseases.* 3. DOI:10.5812/pedinfect.22368.

Martínez Pérez, B., De La Torre Diez, I., López Coronado, M., Sainz De Abajo, B., Robles Viejo, M. & García Gómez, J.M. 2014. Mobile clinical decision support systems and applications: a literature and commercial review. *Journal of Medical Systems.* 38(4). DOI:10.1007/s10916-013-0004-y.

Meyer, S.N., Hougen, A. & Edwards, P. 1951. Experimental error in the determination of tuberculin sensitivity. *Public Health Reports (1896-1970).* 66(18):561-569. DOI:10.2307/4587713.

Moghaddam, R.F. & Cheriet, M. 2012. AdOtsu: An adaptive and parameterless generalization of Otsu's method for document image binarization. *Pattern Recognition.* 45(6):2419-2431. DOI:10.1016/j.patcog.2011.12.013.

Morán-Mendoza, O., Marion, A.S., Elwood, K., D.M., P. & J.M., F. 2007. Tuberculin skin test size and risk of tuberculosis development: a large population-based study in contacts. *The International Journal of Tuberculosis and Lung Disease.* 11(9):1014-1020.

Naraghi, S. 2018. Mobile phone-based evaluation of talent tuberculosis infection. Master's Thesis University of Cape Town.

Naraghi, S., Mutsvangwa, T., Goliath, R., Rangaka, M.X. & Douglas, T.S. 2018. Mobile phone-based evaluation of latent tuberculosis infection: Proof of concept for an integrated image capture and analysis system. *Computers in Biology and Medicine.* 98:76-84. DOI:10.1016/j.compbiomed.2018.05.009.

Otsu, N. 1979. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics* 9(1):62 - 66. DOI:10.1109/TSMC.1979.4310076.

Pooran, A., Booth, H., Miller, R.F., Scott, G., Badri, M., Huggett, J.F., Rook, G., Zumla, A. et al. 2010. Different screening strategies (single or dual) for the diagnosis of suspected latent tuberculosis: a cost effectiveness analysis. *BMC Pulmonary Medicine*. 10(7). DOI: 10.1186/1471-2466-10-7

Radford, A., Metz, L., & Chintala, S. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434.

Risser, N.L., Belcher, D.W., Bushyhead, J.B. & Sullivan, B.M. 1985. The accuracy of tuberculin skin tests: self-assessment by adult outpatients. *Public Health Rep*. 100(4):439–445.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. & Chen, X. 2016. Improved Techniques for Training GANs. *Advances in neural information processing systems*. 2234-2242. arXiv preprint arXiv:1606.03498v1

Siroka, A., Law, I., Macinko, J., Floyd, K., Banda, P., Hoa, B., Tsolmon, B. et al. 2016. The effect of household poverty on tuberculosis. *The International Journal of Tuberculosis and Lung Disease.* 20(12):1603-1609. DOI: 10.5588/ijtld.16.0386

Stork, C., Calandro, E. & Gillwald, A. 2013. Internet going mobile: internet access and use in 11 African countries. *Info.* 15(5):34-51. DOI:10.1108/info-05-2013-0026.

Sugawara, Y., Shiota, S. & Kiya, H. 2018. Convolutional Neural Networks Without Any Checkerboard Artifacts. *26th European Signal Processing Conference*. 1317-1321.

Tajbakhsh, N., Shin, J., Gurudu, S., Hurst, R., Kendall, C., Gotway, M. & Liang, J. 2016. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions On Medical Imaging.* 35(5):1299-1312. DOI:10.1109/TMI.2016.2535302.

Trajman, A., Steffen, R.E., Menzies, D. & Golub, J. 2013. Interferon-Gamma Release Assays versus Tuberculin Skin Testing for the Diagnosis of Latent Tuberculosis Infection: An Overview of the Evidence. *Pulmonary Medicine.* 2013. DOI:10.1155/2013/601737.

Tschandl, P., Rosendahl, C. & Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data.* 5. DOI:10.1038/sdata.2018.161.

UNSDG. 2018. *United Nations Sustainable Development Goals: 3 Good Health and Well-being.*

WHO. 2015. The end TB strategy (No. WHO/HTM/TB/2015.19). *WHO.*

WHO. 2017. Global tuberculosis report 2017. *WHO.*

Wong, S.C., Gatt, A., Stamatescu, V. & McDonnell, M.D. 2016. Understanding Data Augmentation for Classification: When to Warp? *2016 International Conference on Digital Image Computing: Techniques and Applications.* 1-6. DOI: 10.1109/DICTA.2016.7797091.

Yamashita, R., Nishio, M., Do, R.K.G. & Togashi, K. 2018. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging.* 9:611–629. DOI:10.1007/s13244-018-0639-9.

Zeiler, M.D. & Fergus, R. 2014. Visualizing and Understanding Convolutional Networks. *Proceedings of Computer Vision And Pattern Recognition.* 8689:818-833. DOI:10.1007/978-3-319-10590-1_53.