Towards fully automated analysis of sputum smear microscopy images

Marios Zachariou

A thesis submitted for the degree of PhD
at the
University of St Andrews

2024

Identifier to use to cite or link to this thesis:

DOI: https://doi.org/10.17630/sta/858

Towards Fully Automated Analysis of Sputum Smear Microscopy Images

Marios Zachariou

University of St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy

at the University of St Andrews

Supervised by: Dr. Derek J Sloan and Dr. Ognjen Arandjelović

November 2023

Abstract

Sputum smear microscopy is used for diagnosis and treatment monitoring of pulmonary tuberculosis (TB). Automation of image analysis can make this technique less laborious and more consistent. This research employs artificial intelligence to improve automation of *Mycobacterium tuberculosis* (*Mtb*) cell detection, bacterial load quantification, and phenotyping from fluorescence microscopy images.

I first introduce a non-learning, computer vision (CV) approach for bacteria detection, employing ridge-based approach using the Hessian matrix to detect ridges of *Mtb* bacteria, complemented by geometric analysis. The effectiveness of this approach is assessed through a custom metric using the Hu moment vector. Results demonstrate lower performance relative to literature metrics, motivating the need for deep learning (DL) to capture bacterial morphology.

Subsequently, I develop an automated pipeline for detection, classification, and counting of bacteria using DL techniques. Firstly, Cycle-GANs transfer labels from labelled to unlabeled fields of view (FOVs). Pre-trained DL models are used for subsequent classification and regression tasks. An ablation study confirms pipeline efficacy, with a count error within 5%.

For downstream analysis, microscopy slides are divided into tiles, each of which is sequentially cropped and magnified. A subsequent filtering stage eliminates non-salient FOVs by applying pre-trained DL models along with a novel method that employs dual convolutional neural network (CNN)-based encoders for feature extraction: one encoder is dedicated to learning bacterial appearance, and the other focuses on bacterial shape, which both precede into a bottleneck of a smaller CNN classifier network. The proposed model outperforms others in accuracy, yields no false positives, and excels across decision thresholds.

Mtb cell lipid content and length may be related to antibiotic tolerance, underscoring the need to locate bacteria within paired FOV images stained with distinct cell identification and lipid detection, and to measure bacterial dimensions. I employ a proposed UNet-like model for precise bacterial localization. By combining CNNs and feature descriptors, my method automates reporting of both lipid content and cell length. Application of the approaches described here may assist clinical TB care and therapeutics research.

Acknowledgements

General acknowledgements

I would like to express my profound gratitude to the Infection Disease Group at the School of Medicine for their invaluable collaboration over the past three years, which has significantly facilitated my research endeavors. I would like to extend special thanks to my clinical supervisor, Dr. Derek J. Sloan, who has been instrumental in providing essential background knowledge and valuable guidance throughout my research. His support has been pivotal in deepening my understanding of microbiology and clarifying the necessary perspectives for this interdisciplinary study. Dr. Sloan's insights have been crucial in establishing clear criteria for the research, aligning it with the needs and viewpoints of the microbiology field including important ground truth labelling. Additionally, I extend my gratitude to Dr. Wilber Sabiiti and Dr. Bariki Mtafya for contributing their fluorescence microscopy dataset, which constituted Image Set 1. My thanks also go to Dr. Evelin Dombay, not only for providing the ground truth labels but also for supplying the two slides that comprised Image Set 2.

I must express my deepest appreciation to my loving parents, Kypros and Giota Zachariou, who have been my steadfast pillars of support throughout my entire PhD journey. I am also grateful to my dear friends Christos, Andri and Sotiroula for their camaraderie and encouragement. Special gratitude is extended to Martin and Neofytos, who were not only friends but also colleagues, and with whom I have greatly enjoyed exchanging knowledge and ideas. Finally, my heartfelt thanks go to my beloved fiancée, Hope Nicolaou, who has been a constant source of mental support, especially during the most challenging times. Her ability to instill strength in me, regardless of the circumstances, has been invaluable, indeed a truly special woman and person of my entire life.

Finally, my deepest gratitude extends to my Computer Science supervisor, mentor and now friend, Dr. Ognjen (Oggie) Arandjelović. His substantial impact on my academic and professional path cannot be overstated. Oggie's unwavering faith in my potential as both an individual and a researcher, even during times of self-doubt, has been a cornerstone to my journey. Oggie's vast, unlimited knowledge, enthusiasm, and experience has been instrumental in shaping my personal and academic growth. Regardless of the time or circumstances, he has consistently been a pillar of support for me, imparting skills such as professional writing that I once thought were beyond my reach. It is for these reasons and many more that I warmly dedicate this thesis to him. I eagerly anticipate future collaborations and new projects, as well as the continuation of our friendship for the remainder of my time.

Funding

This work was supported by a Wellcome Trust Institutional Strategic Support Fund award to the University of St Andrews, grant code 204821/Z/16/Z.

Research Data/Digital Outputs access statement

Research data underpinning this thesis are available at `https://doi.org/10.17630/28cc5ee6-7cfa-4335-b591-e1e9247dc4f6`

Declaration

Candidate's Declarations

I, Marios Zachariou, hereby certify that this thesis, which is approximately 44000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

I was admitted as a research student and as a candidate for the degree of Doctor of Philosophy in October 2019; the higher study for which this is a record was carried out in the University of St Andrews between 2019 and 2024.

Date: 15/11/2023

Signature of candidate:

Supervisor's Declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date: 15//11/2023

Signature of supervisor(s):

Permission for Electronic Publication

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use unless exempt by award of an embargo as requested below, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration, or have requested the appropriate embargo below.

The following is an agreed request by candidate and supervisor regarding the electronic publication of this thesis:

> *Access to printed copy and electronic publication of thesis through the University of St Andrews.*

> **Printed copy**
>
> No embargo on print copy

> **Electronic copy**
>
> No embargo on electronic copy

I, Marios Zachariou, confirm that my thesis does not contain any third-party material that requires copyright clearance.

Date: 15/11/2023

Signature of candidate:

Signature of supervisor(s):

Underpinning Research Data or Digital outputs

Candidate's declaration

I, Marios Zachariou, understand that by declaring that I have original research data or digital outputs, I should make every effort in meeting the University's and research funders' requirements on the deposit and sharing of research data or research digital outputs.

Date: 15/11/2023

Signature of candidate

Permission for publication of underpinning research data or digital outputs

We understand that for any original research data or digital outputs which are deposited, we are giving permission for them to be made available for use in accordance with the requirements of the University and research funders, for the time being in force.

We also understand that the title and the description will be published, and that the underpinning research data or digital outputs will be electronically accessible for use in accordance with the license specified at the point of deposit, unless exempt by award of an embargo as requested below.

The following is an agreed request by candidate and supervisor regarding the publication of underpinning research data or digital outputs:

No embargo on underpinning research data or digital outputs.

Date: 15/11/2023 Signature of candidate

Date: 15/11/2023 Signature of supervisor(s)

Contents

List of Figures

List of Tables

Abbreviations

R^2 coefficient of determination.

Mtb *Mycobacterium tuberculosis*.

AFB Acid-Fast Bacteria.

AI artificial intelligence.

AUC area under the ROC curve.

BCG Bacillus Calmette-Guérin.

CAT channel area thresholding.

CNN convolutional neural network.

CT computed tomography.

CV computer vision.

CXRs Chest radiographs.

Cycle-GANs cycle-consistent generative adversarial networks.

DCNN deep convolutional neural network.

DET decision error trade-off.

DFT discrete Fourier transform.

DL deep learning.

DNA deoxyribonucleic acid.

DST drug susceptibility testing.

Faster-RCNN Faster Region-based convolutional neural networks.

FD Fourier descriptors.

FHDT fuzzy and hyco-entropy-based decision tree.

FN false negative.

FOV field of view.

FP false positive.

FPR false positive rate.

GA-NN genetic algorithm neural network.

GANs Generative adversarial networks.

GB gigabytes.

GPU graphical processing unit.

HEDT hyco-entropy-based decision tree.

HMLP hybrid multi layered perceptron.

HOG histogram of gradients.

KD knowledge-distillation.

LED light-emitting diode.

LMICs Low and Middle Income Countries.

LP Lipid poor.

LR Lipid rich.

LTBI Latent Tuberculosis Infection.

LTR LipidTox Red.

MAE mean absolute error.

MAPE mean absolute percentage error.

MDG Millennium Development Goals.

MDR-TB Multidrug-resistant tuberculosis.

ML machine learning.

MLP Multi-layer perceptron.

MSE mean squared error.

MSVR multi-output support vector regressor.

MTC Mycobacterium tuberculosis complex.

MWI Modified William index.

NLP natural language processing.

NN Neural network.

NRPs Non-Replicating Persisters.

PCR polymerase chain reaction.

PHIL CDC Public Health Image Library.

PPR predictive positive rate.

RBF radial basis function.

RF Random Forest.

RGB red-green-blue.

RMSE root-mean square error.

ROC receiver operating characteristics.

SD Sorensen-dice.

SIFT scale invariant feature transform.

SOP standard operating procedures.

SURF speeded up robust feature.

SVM support vector machine.

SVNN support vector neural network.

TB tuberculosis.

TB-AI Tuberculosis - artificial intelligence.

TN true negative.

TNR true negative rate.

TP true positive.

TPR true positive rate.

ViT Vision Transformer.

WHO World Health Organisation.

WSI whole slide image.

XAI explainable artificial intelligence.

ZNSM–iDB Ziehl–Neelsen sputum smear microscopy image database.

Chapter One

Prologue

This thesis establishes on an interdisciplinary nexus at the confluence of computer science and medicine, aiming to advance the domain of tuberculosis (TB) diagnosis and treatment monitoring through the innovative application of automated image analysis and machine learning methodologies. The menace of TB remains a global health challenge, with the diagnosis and monitoring processes being critical yet intricate components in the fight against the disease. The goal of this research is to harness the capabilities of fluorescence microscopy, combined with the analytical power of machine learning, to develop an automated system capable of accurately diagnosing TB and monitoring treatment efficacy.

The objectives outlined herein focus on the design, development, and validation of machine learning algorithms tailored for the analysis of fluorescence microscopy images. These objectives are twofold: first, to achieve a level of diagnostic accuracy that meets or exceeds current standards; and second, to provide a means of monitoring treatment progress, thereby offering a potentially transformative tool for healthcare professionals. In pursuit of these objectives, this work has yielded several publications, detailing the methodologies employed, the datasets curated, and the experimental results obtained, thereby contributing to the body of knowledge in both the medical and computer science communities.

Motivation for this study stems from the critical need to improve TB diagnostic and monitoring tools to accelerate treatment initiation, enhance patient outcomes, and reduce transmission. The integration of machine learning with fluorescence microscopy represents a novel approach in this regard, offering the promise of greater efficiency, accuracy, and accessibility in TB healthcare services.

Before delving into the technical intricacies of this research, it is imperative to establish a foundation accessible to readers across disciplines. Chapter 1 will therefore provide the necessary medical background on TB, its global impact, and the role of diagnostics and treatment monitoring in managing the disease. This introductory chapter ensures that readers, regardless of their primary field of study, possess a sufficient understanding of the medical context, enabling a comprehensive grasp of the subsequent chapters that detail the research's technical aspects.

In summary, this thesis presents a cohesive exploration of the potential of machine learning in revolutionizing TB diagnosis and treatment monitoring through automated image analysis of fluorescence microscopy images. By laying out the goals, objectives, and motivations, this prologue sets the stage for a detailed exposition of interdisciplinary research aimed at bridging the gap between technological innovation and medical application.

1.1 Using artificial intelligence to improve TB control

Sections 2.1 to 2.3 of Chapter 2 will advocate that innovative approaches to diagnosis and treatment are required to regain lost ground in efforts to meet international TB control targets. Sections 2.6 and 2.7 have illustrated the strengths and limitations of existing tools, emphasising elements where slow manual procedures are a bottleneck to progress. In recent decades, the field of artificial intelligence (AI) has expanded rapidly. Computer-based machine and deep learning (DL) algorithms are being developed to perform many activities previously done by humans including medical diagnostics [101, 271] and biomedical research on infectious diseases [252]. Within the sphere of TB, there has been progress in using AI methods (some researchers in the field also refer to it as Tuberculosis - artificial intelligence (TB-AI)) tools to screen new compound databases and model structure-activity relationships of novel chemical entities during drug discovery [250]. Advanced image analysis tools have also been developed for computer assisted interpretation of radiological tests (CXRs and CT scans) in presumptive TB patients [40, 127] but progress in automating sputum smear microscopy lags further behind.

1.1.1 The rationale for computer-based automaton of TB microscopy

Slower progress in automation of TB microscopy could reflect several factors. The complexity of resolving this issue arises in part from the challenges associated with segmenting and tracking *Mtb* cells, which often exhibit irregular shapes and tend to aggregate, as noted by Chung *et al.* [39]. A more comprehensive discussion in relation to these factors will be presented later throughout all the chapters of this thesis. There may also be an underlying assumption that smear microscopy is soon to be replaced for primary TB diagnosis by newer molecular methods, which disincentivises scientific effort to improve it. However, I contend that further investigation of AI approaches to sputum smear microscopy is important for three reasons. Firstly, microscopy is still widely used and there are practical obstacles to the implementation of replacement techniques in many settings as outlined in Section 2.6. Secondly, microscopy still has a role, not yet supplanted by any other method, in treatment monitoring, as outlined in Section 2.7.2. Thirdly, and perhaps most importantly, microscopy may have specific research value in understanding heterogeneity in TB treatment response at the level of individual cell morphology [39, 239] as will be outlined in Section 2.7.3. There is no replacement technology for that function in the foreseeable future. There are clear examples where detailed microscopy-based study of single cells have led to important advances in our understanding of crucial questions in pathogens which cause human infection. Multicolour fluorescence microscopy has contributed to elucidation of the developmental morphologies of the malaria parasite Plasmodium falciparum [10]. High content confocal microscopy imaging has been successfully performed to help identify factors which influence disease severity in infections caused by *M abscessus*, an organism in the same bacterial family as *Mtb* [23]. These results illustrate potential applications for similar tools in TB.

There are significant barriers to efficiently using microscopy for clinical case management and clinical research on *Mtb* without automation. These principally relate to the time investment, operator-dependency, and subjectivity. Skilled diagnostic TB microscopists can only manage a maximal workload of 20-25 specimens per day [185]. Obtaining and analysing images for academic evaluation of single cells is even more challenging; the laboratory practitioner must stain and examine the slide, detect relevant FOVs at the correct magnification, and take digital photographs of them for subsequent use. Digital images of complete microscopy slides feature resolutions that extend into the hundreds of thousands of pixels and result in files with considerable sizes, approximately 19 gigabytes (GB). However, the vast majority of these images encompass FOVs that are devoid of bacteria. As a result, microscopists often find themselves expending considerable time scouring for and capturing fewer than 50–150 selected salient FOVs. These cropped images usually adopt dimensions of 2000 × 1000 pixels. Some slides from clinical samples are especially challenging because AFB might have odd appearances and non-bacterial artefacts inside the sputum matrix can mimic *Mtb*. Overall, therefore there is a strong case to develop AI approaches to TB microscopy which will improve consistency and objectivity of results, and increase throughput for clinical diagnosis, treatment monitoring and academic research.

1.2 Aims and objectives of the work described in this thesis

The primary objective of the research detailed in Chapters 3– 7 involves innovating and developing methods for advancing the automation of TB microscopy through AI techniques. Specific objectives in pursuit of this aim were to:

- Review the literature on existing knowledge within the field as well as its limitations. This is the basis of Chapter 3, elements of which have already progressed to peer-reviewed publication [266].

- Examine FOVs previously imaged from a clinical dataset of TB microscopy images, using geometry-based and deep learning approaches, in order to automate detection and counting of *Mtb* bacteria. This is the basis of Chapter 4 and 5, elements of which have already progressed to peer-reviewed publication [265]

- Develop a method to extract and classify salient FOVs from whole microscopy slides containing *Mtb* bacteria, in order to automate the process of cropping areas of interest from the sample. This is the basis of Chapter 6, elements of which have already progressed to peer-reviewed publication [263].

- Develop a method to automate examination of *Mtb* bacteria within cropped FOVs on images from a clinical dataset of TB microscopy images, in order to report cell length and lipid content. This is the basis of Chapter 7, elements of which have already been presented as a conference paper [264], and progressed to a peer-reviewed publication [262].

Chapter Two

Biomedical background

Chapter Abstract – This Chapter provides background information on the clinical and microbiological problem which the work of this thesis seeks to address. It describes the epidemiology, pathophysiology, clinical characteristics, current diagnostic tests, and treatment of TB. It pays particular attention to the benefits and drawbacks of microscopy as a diagnostic tool, in order to outline the potential advantages of using artificial intelligence to automate this procedure within the framework of this field of work.

2.1 Historical overview

The disease tuberculosis (TB) has afflicted humankind for millennia [64]. It remains one of the top ten causes of death worldwide, causing approximately 5000 fatalities each day. The bacterium, *Mycobacterium tuberculosis* (*Mtb*), is the main micro-organism which causes clinical TB in humans and was first identified by Robert Koch in 1882 [54]. At that time up 25% of all deaths in Europe were attributed to the disease [121]. Throughout history, TB has been strongly linked to poverty and adverse social circumstances. Deaths from TB in industrialised countries decreased in the first half of the 20th Century, largely due to improvements in housing, nutrition, and personal income [244]. Advances in drug development between the 1940s and the 1980s culminated in establishment of curative antibiotic combination regimens. For a brief period, global TB control looked achievable. However, in the early 1990s, the collapse of Public Health infrastructure in former Soviet states heralded a new epidemic of Multidrug-resistant tuberculosis (MDR-TB) TB in Eastern Europe [231]. Around the same time, the Human Immunodeficiency Virus (HIV) pandemic spurred an upsurge in all forms of TB worldwide, particularly in southern Africa [46]. In 1994, the World Health Organisation (WHO) declared that TB was a “global emergency” [160]. This prompted a series of strategic public health initiatives.

In 2000, Target 6c of the United Nations Millennium Development Goals (MDG) set out to reverse the rising incidence and reduce deaths from TB by 50% by 2015 [236]. These goals were attained, so a new "End TB strategy" was unveiled with the ambitious vision of ending the worldwide TB epidemic by 2030 [162]. In 2016, the global rate of decline in people falling sick with TB was 1.5% per year. Considerable investment, and progress in research and innovation to improve TB diagnosis and treatment, was always going to be necessary to accelerate progress and achieve these new goals (see Figure 2.1) [162]. Unfortunately, the unexpected onset of the Coronavirus-19 (Covid-19) pandemic derailed many planned activities and reversed many of the successes of recent years [59]. There is an urgent need to regain public health and scientific momentum now.

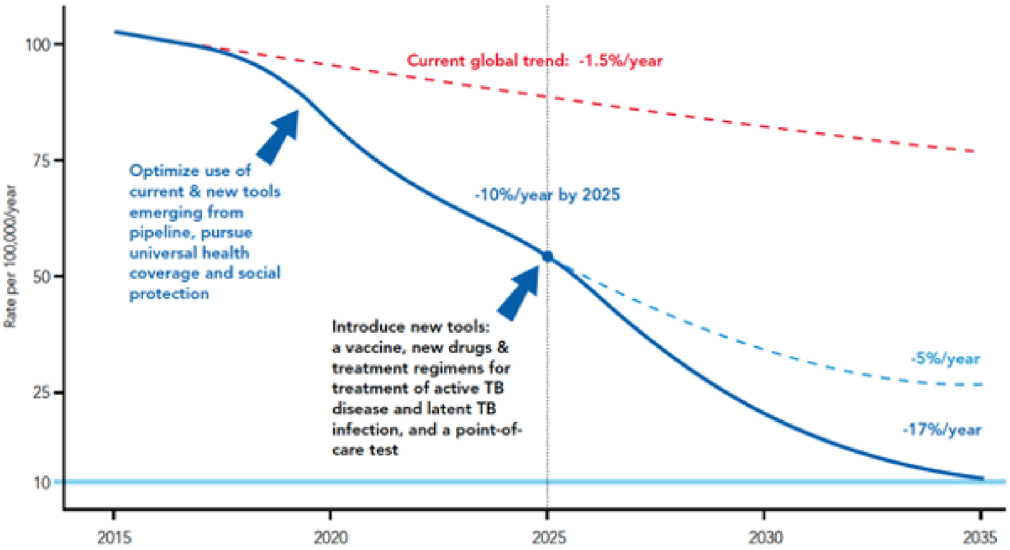


Figure 2.1: Research and innovation requirements of "End TB Strategy"

2.2 Current global epidemiology of TB

An unusual feature of TB is that most people who become infected with the causative organism do not become ill. This phenomenon is called Latent Tuberculosis Infection (LTBI) and is estimated to affect 23% of the world's population (1.7 billion individuals) [102]. 5-10% of people with LTBI progress to suffer from active TB disease over the course of their lives. Approximately 10 million people fall ill with TB disease every year; this figure has remained relatively constant for 20 years, but the estimated burden rose by 4.5% from 10.1 million to 10.6 million between 2020 and 2021 [165]. From 2005 to 2019, global mortality from TB decreased from 1.7 million to 1.2

million deaths per year [161, 163]. However, this trend reversed in 2020 (1.3 million deaths in HIV-negative, and 214,000 deaths in HIV positive people respectively) and 2021 (1.4 million deaths in HIV-negative and 187,000 deaths in HIV-positive people respectively) [165, 164]. These recent data provide a stark illustration of the damaging effect of Covid-19 on TB control. The worldwide burden of TB does not fall evenly. Most people who develop active disease live in WHO regions of South-East Asia (45%), Africa (23%) and the Western Pacific (18%). In 2021, 30 high burden countries accounted for 87% of incident TB cases globally, and 8 countries accounted for more than two-thirds of the total (see Figure 2.2): India (28%), Indonesia (9.2%), China (7.4%), the Phillipines (70%), Pakistan (5.8%), Nigeria (4.4%), Bangladesh (3.6%), and the Demographic Republic of the Congo (2.9%) [165].

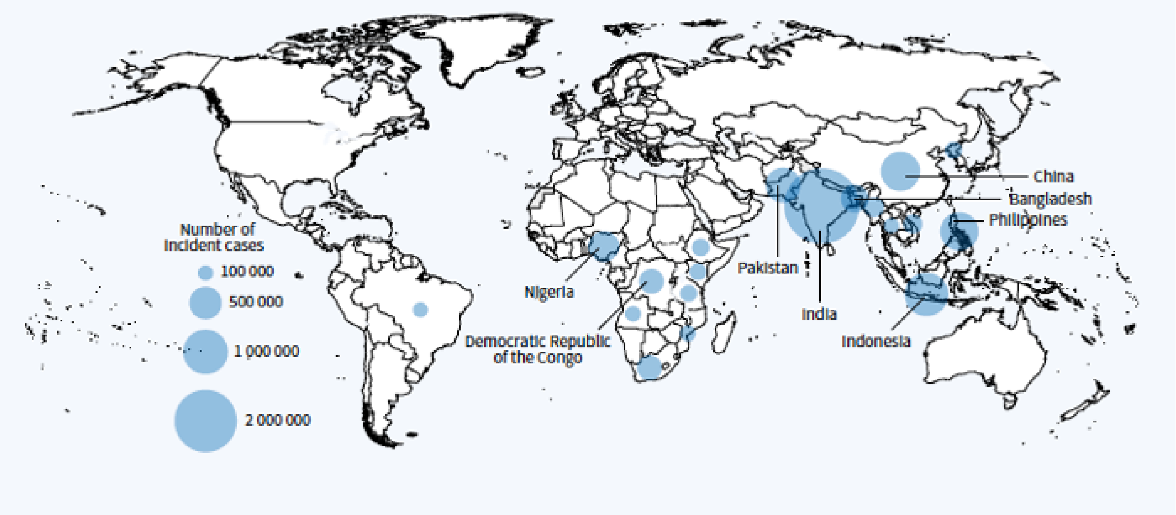


Figure 2.2: TB incidence in 2021; countries with at least 100 000 incident cases

When reviewing these epidemiology data, it should be remembered that the figures are estimates based on mathematical modelling of surveillance studies. Of the estimated 10 million new incident TB cases each year only 6-7 million are actually detected and notified to healthcare authorities [191]. Closing this detection gap is essential and requires improvements to public health strategies and the technical tools used for TB diagnosis. Whilst TB is generally considered to be an antibiotic-curable bacterial infection, the last decade has seen growing concern about antimicrobial resistance. At least 500,000 people/year with active TB (5% of the total number of new TB cases) are now believed to be infected with *Mtb* strains which have developed genetic mutations conferring resistance to rifampicin, the most important first-line drug used in current treatment combinations [165]. A growing number of complex resistance patterns, involving *Mtb* strains which cannot be treated with other important drugs, have further resulted in categorisation of ‘multi-drug’ or ‘extensively-drug’ resistant TB. International surveillance is ongoing to monitor the emergence of drug-resistance, which is considered a major threat to long-term TB control.

2.3 TB control: current priorities and goals

As described in Section 2.1, WHO is currently attempting to deliver a strategy to end the global TB epidemic [162]. Specific indicators of this are to achieve:

- 95% reduction by 2035 in the number of TB deaths compared with 2015
- 90% reduction by 2023 in new TB cases compared with 2015
- Zero TB-affected families facing catastrophic costs due to TB by 2035.

To be on target for these goals, the first milestone planned for a 35% drop in new TB cases between 2015 and 2020. However, the net decline over that period was only 5.9% [165]. Although Covid-19 is implicated in missing this milestone, it was not the only factor because progress was behind schedule even in the period from 2015 to 2019. Wherever possible, the most effective biomedical tool to curtail infectious disease epidemics is vaccination. For TB, the only currently available vaccine is Bacillus Calmette-Guérin (BCG), which has existed for over 80 years and is only partially effective. About 20% of those who receive the vaccination are protected against contracting infection, while advancement of disease after acquiring infection is halted in 60%. Duration of protection varies by population and geographical setting but some benefit typically seems to persist for 15 to 20 years [216]. Recent research into development of better vaccines has identified promising candidates [223] but it will be some time until these are ready for widespread use. Given the limitations of the currently available vaccine, TB control presently relies heavily on a ‘detect and treat disease’ approach. Prompt initiation and successful completion of effective drug therapy not only cures individual patients, restoring their quality of life and reducing mortality rates, but it prevents onward *Mtb* transmission and reduces the risk that antibiotic-resistant strains of bacteria will emerge [217]. The commonest antibiotic treatment regimen for TB takes 6 months to complete [167]. Optimal deployment of a ‘detect and treat’ strategy for TB control requires better diagnostic tools, and advances in drug development to shorten treatment duration. Improving awareness amongst communities, public and private health care providers, politicians, and funding bodies is also pivotal to securing adequate long-term resources for sustained TB control.

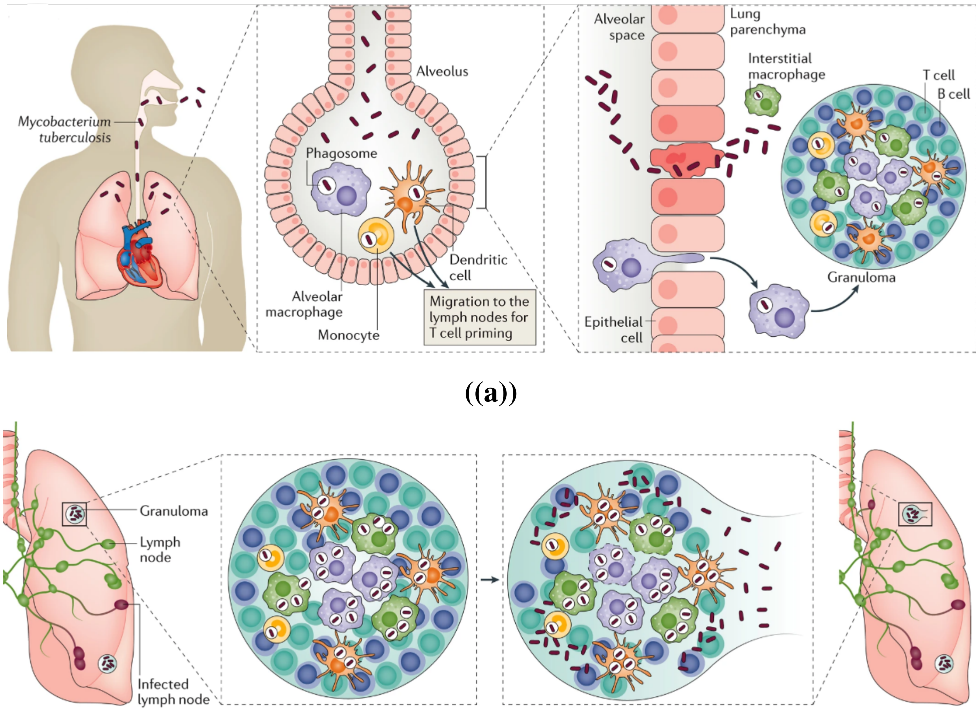
2.4 Transmission of TB

Although *Mtb* is the main causative organism, several other genetically related *mycobacteria* can also lead to TB. Collectively these are referred to as the Mycobacterium tuberculosis complex (MTC) and other members include: *M africanum*, *M bovis*, *M canetti* and *M microti* [106]. As most of these organisms are >99% genetically identical to *Mtb*, they are regarded as microbiological synonyms [186]. The commonest anatomical site of human disease is the lungs (pulmonary TB), and the principal route of transmission is by respiratory droplets or aerosols. When a person with active pulmonary TB coughs or sneezes they expel infectious particles 0.5–5.0μm in diameter into the air [230, 275, 235]. A single cough can release up to 40,000 particles, each one of which may transmit the disease because inhalation of fewer than 10 bacteria may be sufficient to establish a new focus on infection [155]. Most newly infected individuals are between 15 and 54 years old, the age group with greatest likelihood of sustained, frequent contact with others. Living in overcrowded, under-ventilated environments also increases transmission, explaining why the highest densities of TB cases occur in settings of urban poverty [213].

Patients who expel higher bacterial concentrations when they cough are more likely to transmit TB to others [168], so diagnostic tests which measure bacterial load in expectorated sputum may be more useful when considering transmission risk than those which only describe ‘positive’ or ‘negative’ *Mtb* detection.

2.5 TB pathogenesis and clinical disease

In most cases, inhaled *Mtb* bacteria travel through the respiratory tract to reach immune cells called alveolar macrophages in the lungs [41]. These macrophages engulf the bacteria and recruit other immune cells including monocytes, dendritic cells and T and B lymphocytes to form pathological lesions called granulomas [229]. In individuals with strongly functioning immune systems, these granulomata may contain the infection without causing any symptoms of illness at all, creating the phenomenon of LTBI which was outlined in Section 2.2. People with LTBI usually do not know that they have encountered *Mtb* and cannot transmit the disease to others. 90-95% of the time they never become unwell from TB. However, in 5-10% of people the immunological balance in these granulomata is disrupted, containment of LTBI infection is lost, and active TB disease follows. This is more likely in scenarios where the person’s immune system is weakened by extremes of age, co-existent other illnesses (most notably HIV infection, but also other chronic health problems like diabetes mellitus or cancer), or use of immune-suppressant medications (e.g., corticosteroids or biological therapies for rheumatological and connective tissue disease). As active TB disease develops, host cells within granulomata break down and a liquid ‘caseous’ core forms within the pathological lesions. *Mtb* bacteria start to replicate more intensely and can spread, both to adjacent lung tissue and to other parts of the body [229]. Figure 2.3 [172] provides a diagrammatic illustration of the pathogenesis of LTBI and active TB infection described above



((a))

((b))

Figure 2.3: (a) Latent TB infection (LTBI): *Mtb* cells enter the respiratory tract and are engulfed by alveolar macrophages which interact with other immune cells in the lung to form granulomas, the hallmark pathological lesion of TB. (b) Active TB disease: When immunological control of LTBI breaks down, *Mtb* starts to replicate, and granulomas can break down releasing bacteria into other parts of the lung and elsewhere in the body.

In 85% of cases, this process is mainly limited to the respiratory tract and causes the symptoms and signs of pulmonary TB. Increased inflammation around caseous granulomata damages healthy tissue and large holes known as cavities can form in the lungs. Patients develop a productive cough (sometimes containing blood), alongside breathlessness and chest pain which can progress to respiratory failure and death if not appropriately treated. Constitutional symptoms of fever, night-sweats, exhaustion, and weight-loss are also well described. In 15% of cases, the same inflammatory process of active TB occurs in other parts of the body. This is known as extrapulmonary TB, and the associated clinical features are determined by anatomical location. For example, if the central nervous system is affected meningitis or paralysis can develop which carries a high mortality rate. Other sites of extrapulmonary TB include peripheral lymph nodes, the bone and spine, gastrointestinal and genitourinary systems, and the external lining of the heart (known as the pericardium).

2.6 TB diagnosis

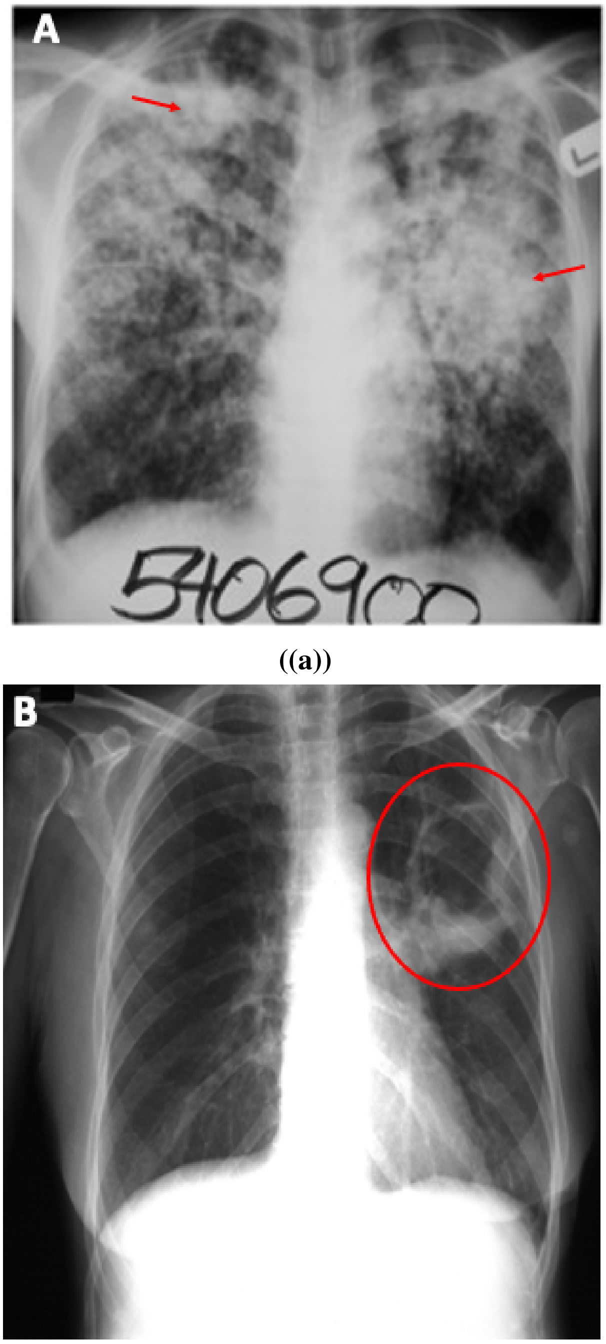
Whilst the clinical consequences of active TB can be dramatic, *Mtb* is a slowly replicating bacteria (doubling time of ~20 hours, compared to ~20 minutes for other pathogenic bacteria such as Escherichia coli) and the disease tends to progress slowly, getting worse over weeks rather than days. This should provide opportunities to diagnose the illness and provide treatment before it becomes severe. Early diagnosis also help prevent transmission [81, 69]. However, as described in Section 2.2, the diagnosis of TB can be difficult, partly because existing tools are sub-optimal. Research to improve them is urgently required.

The specific diagnostic methods used for active TB depend on the clinical presentation and site of disease so are different for each patient. In general, they comprise microbiological techniques (which seek to identify and characterise the infecting bacteria) and radiological techniques (using radiographs and computed tomography (CT) scans to identify typical patterns of tissue damage in affected organs). Chest radiographs (CXRs) from patients with pulmonary TB are shown in Figure 2.4 to exemplify the role of this tool and to underline the pathological processes in the lungs outlined in Section 2.5.

As 85% of active TB cases are pulmonary and the dataset used for most of the work described in this thesis comprises microscopy images obtained from sputum, this remainder of this section will focus on microbiological approaches for the diagnosis of pulmonary TB. An important advantage of microbiological methods is that the only way to really know that *Mtb* is present, is to directly detect it, and microbiology tools are the only means to do that. These tools can be subdivided into the traditional approaches of smear microscopy and *mycobacterial* culture, and more recently adopted modern molecular techniques.

2.6.1 Smear microscopy

Traditionally, sputum smear microscopy has been the predominant approach for diagnosing pulmonary TB. It remains the most frequently used method in many Low and Middle Income Countries (LMICs) [172]. Smear microscopy has several benefits, including the low cost of the laboratory equipment required and the fact that the test can often be done in primary healthcare clinics or mobile laboratories close to the patient, providing results in a short amount of time (ideally on the same day as sample collection).



((a))

((b))

Figure 2.4: Anonymised CXRs images provided by Dr D Sloan (consent for use in place: the number on the bottom of panel of (a) is not a patient identifier). (a) extensive inflammation in both lung fields: ‘normal’ lung tissue is ‘black’ on CXRs, so all fluffy white shadowing, indicated by red arrows is diseased lung. (b) cavitation occurs when a large section of lung tissue is completely destroyed and only a ‘hole’ remains, indicated within the red circle. Cavities contain very large numbers of bacteria.

The principle underpinning microscopy-based diagnosis of TB is that *mycobacteria* cells, including *Mtb*, have very thick, waxy, lipid-rich cell walls which take up specific dyes and then resist decolouration with dilute acid rinse. The result is that dye is concentrated in short rod like structures, approximately 0.2-0.5×1.0-7 μm in size, called Acid-Fast Bacteria (AFB) [230, 275, 235]. Bacterial species with less complex cell walls cannot retain the same dyes so when appropriately stained biological samples are viewed down a microscope, the AFB are selectively identifiable. Two main staining procedures are used for AFB identification: the Ziehl-Neelsen method which requires a light microscope and the Auramine O method which requires a fluorescence microscope [275].

There are also disadvantages of smear microscopy. Firstly, whilst the financial expense of equipment for it is low, the cost in terms of staff salaries and time can be high. For example, in the early 2000s, a district diagnostic laboratory in Malawi observed that 43% of total staff workload was allocated to TB microscopy. In addition to making TB diagnosis more difficult and increasing the time between sample collection and results, this also adversely affects the limited capacity of skilled personnel to do other equally important work (e.g., diagnostic tests for malaria, cross-matching blood for transfusion and other high priority diseases) [151].

Secondly, the reliability of smear microscopy results is highly dependent on the skill of the microscopist reading the stained slides. All medical diagnostic procedures based on image analysis contain an inherent degree of subjectivity, especially when the images are interpreted manually. Investigators from Vietnam have previously reported that human error can be implicated in mis- or late diagnosis of TB by microscopy [133]. Some initiatives (e.g., auto-focusing microscopes, not to be confused with auto-focusing algorithms that we will see later) have been considered to try and tackle this problem [49]. These issues highlight a question which is motivates this thesis: could automation of microscopy slide reading and image analysis accelerate, standardise, and improve the quality of sputum smear microscopy procedure?

Finally, an unavoidable problem with smear microscopy cannot provide all the desired information about the AFB which are seen. *Mtb* and other MTC organisms are the most important disease-causing AFB but they are not the only micro-organisms with these staining characteristics so it is occasionally possible that a correctly interpreted microscopy slide might still result in the wrong diagnosis. Drug-susceptible and drug-resistant bacteria are indistinguishable at microscopy, so additional investigations are necessary to identify the TB patients who cannot be routinely treated with first-line antibiotics and require more specialist care.

2.6.1.1 Light microscopy (Ziehl-Neelsen staining)

Standardised protocols for the main TB smear microscopy methods have been published by international agencies [175]. As all the data within this thesis are derived from electronic images of smear microscopy of sputum provided by TB patients, key steps of the laboratory methods are provided here to provide background on the process by which the data were generated.

Prior to smear microscopy by any method, sputum expectorated by presumptive pulmonary TB patients (anyone who is coughing and for whom TB is a possible diagnosis) is mechanically and chemically homogenised to ensure that any bacteria present are evenly distributed through the sample. 10-20μl aliquots are heat-fixed onto glass slides, which immobilises them for staining and kills most of the infectious material, increasing safety for laboratory workers. Heat-fixed slides are colourised by flooding with kinyoun carbol fuschin and allowed to dry for 5 minutes before washing with distilled water. They are then de-colourised with 3% acid-alcohol before repeat washing with distilled water. Finally, methylene blue counter stain is added, left for 2 minutes, and washed off in the same way. Slides are then left to air-dry and viewed down a light microscope at ×1000 magnification (usually ×100 from the objective lens and ×10 from the ocular eyepiece lens). They should be viewed systemically, with three horizontal sweeps across the length of the slide before being reported (Figure 2.5).

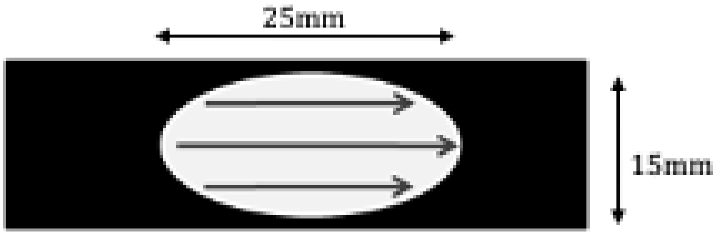


Figure 2.5: 10-20μl homogenised sputum should be smeared onto slides aiming for maximum smear diameter of 25mm ×15mm and keeping the smear thin for ease of viewing (multiple cell layer thick slides are hard to read). During microscopy three horizontal sweeps (direction shown by arrows) should be made of each smear for optimal detection and count of bacteria

If AFB are present, they are visualised as red rods whilst host cell debris in the background is light blue (Figure 2.6) [275]. For patients in whom pulmonary TB is diagnosed by smear microscopy the bacterial load of *Mtb* cells in the sputum can be semi quantified by counting the cells visible across a representative range of fields of view (FOVs). Standardised systems have been developed for this by the International Union Against Tuberculosis and Lung Disease, as shown in Table 2.1 [175]. Grading of the bacterial load in this way can be important, as higher concentrations of *Mtb* in sputum are associated both with increased risk of disease transmission to other people [168] and increased risk that antibiotic treatment will be unsuccessful [214].

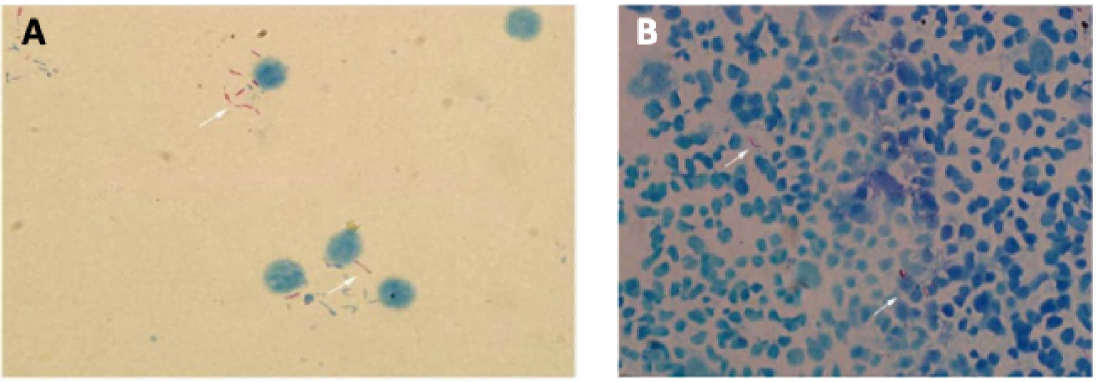


Figure 2.6: AFB are labelled red (shown by white arrows in both images). The amount of cellular debris, stained blue, is much more in B than A, illustrating the heterogeneity which can occur during the staining process which can increase difficulty and variability of slide reading.

Table 2.1: (a) At least 5 minutes should be taken to read 100 FOVs before reporting the slide as negative. (b) A finding of 1-3 bacilli in 100 FOVs does not correlate well with culture positivity. It is recommended that a new smear be prepared from the same sputum specimen and be re-examined. (c) In practice most microscopists read a few FOVs and confirm the findings by a quick visual scan of the remaining FOVs

AFB Counts	**Grading**
No AFB in at least 100 microscopy FOVs[a]	0 or ‘negative’
1-9 AFB in 100 FOVs[b]	‘Scanty’; record actual number of AFB counted
10-99 AFB in 100 FOVs	1+ *or* ‘+’
1-10 AFB per field in at least 50 FOVs[c]	2+ *or* ‘++’
>10 AFB per field in at least 20 FOVs[c]	3+ *or* ‘+++’

2.6.1.2 Fluorescence microscopy (Auramine O staining)

Even though fluorescence microscopy was first developed in 1930, it was not until 2008 that it was widely used as a diagnostic tool for TB. Sputum smear examination adheres to the same principles as light microscopy. However, an acid-fast fluorochrome dye (Auramine O) is used instead of kinyoun carbol fuschin to colour the slides, a lower concentration of acid-alcohol (0.5%) is used for decolourisation and 0.5% potassium permanganate is used as a counter-stain. A fluorescence microscope, with a bright light source, either halogen or light-emitting diode (LED), and appropriate wavelength filters (Auramine O peak λexcitation>432nμ, λemission>499nμ) is used to view the slides at ×200-×400 magnification (usually ×20-×40 from the objective lens and ×10 from the ocular eyepiece lens). Slides should be viewed with 24 hours of staining, or the fluorescence can fade. If AFB are present, they are visualised as bright yellow/green rods against a black background (see Figure 2.7) [175].

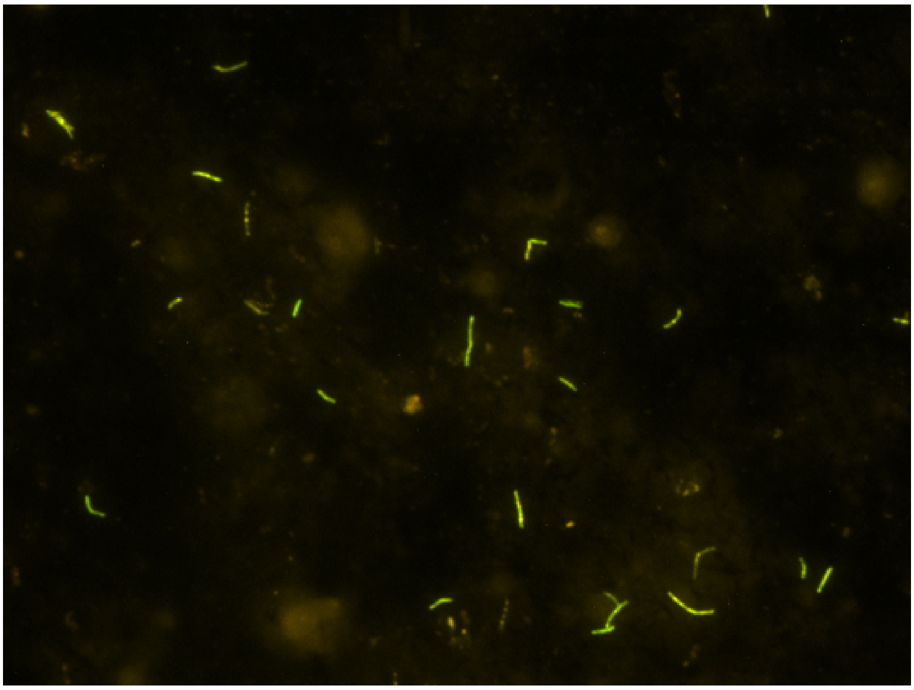


Figure 2.7: Typical sputum smear microscopy images using Auramine O staining. Microscopy image provided by Dr D. Sloan.

Auramine O stained sputum smears viewed by fluorescence microscopy can also be semi quantitatively graded for bacterial load. Different grading scales are used than for Ziehl-Neelsen, considering the different magnification ranges used (Table 2.2) [175].

Many studies have shown consistently better performance for fluorescence compared to conventional light microscopy for TB diagnosis. In a meta-analysis of sensitivity and specificity for the two techniques compared to a gold standard of *mycobacterial* culture, light microscopy with Ziehl-Neelsen staining had diagnostic sensitivity (denoting ability to detect all true positive samples containing AFB) ranging from 0.34 to 0.94, whilst fluorescence microscopy with auramine-based staining had sensitivity ranging from 0.52 to 0.97 [214]. The wide ranges serve

Table 2.2: (a) Confirmation required by another technician, or prepare another slide before reporting this result

What you see (×200)	What you see (×400)	What to report
No AFB in one length	No AFB in one length	No AFB observed
1-4 AFB in one length	1-2 AFB in one length	Confirmation required[a]
5-49 AFB in one length	3-24 AFB in one length	Scanty
3-24 AFB in one field	1-6 AFB in one field	1+ *or* '+'
25-250 AFB in one field	7-60 AFB in one field	2+ *or* '++'
250 AFB in one field	60 AFB in. one field	3+ *or* '+++'

as a reminder of how user-dependent and subjective smear microscopy can be. Even allowing for this the increase in sensitivity on fluorescence microscopy is notable. The primary reason is that AFB, such as *Mtb*, are easier to detect using fluorescence, since the colours increase the contrast between the bacteria and their surroundings.

Whilst concern has been raised that non-organic artefactual matter in the sputum matrix can also fluorescence brightly on auramine-stained slides, specificity (denoting ability to avoid false positive samples where AFB were incorrectly identified) was 0.99 for both light fluorescence microscopy methods [214]. Following investment and innovation by the Foundation for Innovative and New Diagnostics (FIND, `https://www.finddx.org`) and others, easy to maintain LED-based fluorescence microscopes (e.g., PrimoStarTM iLED) have become more affordable [5]. Consequently, light microscopy is being phased out and replaced with fluorescence tools in many countries.

In summary, sputum smear microscopy offers diagnostic information rapidly. When performed correctly, it offers a high specificity for detecting *Mtb* cells. Although microscopy laboratory supplies are inexpensive, the process is time-intensive, which impacts laboratory personnel expenses. The wide ranges of diagnostic sensitivity reflect the complexity and subjectivity of the procedure.

2.6.2 Mycobacterial culture

Whilst sputum smear microscopy has long been the most commonly used tool for detection of pulmonary TB, the ability to grow *Mtb* in *mycobacterial* culture remains the gold standard for diagnostic confirmation. Various techniques are available to do this, using both solid and liquid culture media [175]. Recovery of fewer viable *Mtb* cells is possible from sputum culture than smear microscopy, i.e. a starting bacterial load of 10^2 bacteria/mL can be detected by some culture-based methods compared to 10^4 bacteria/ml by microscopy [240] and the isolates obtained can be used for drug susceptibility testing (DST) to confirm the presence or absence of antibiotic resistance [175], so the advantages are clear.

However, in practical terms there are major challenges to culture-based diagnostics too. Firstly, as *Mtb* is a slowly growing bacteria it can take 2-6 weeks to generate a result [139] and clinical decision-making needs to move faster than that. Secondly, *mycobacterial* culture grows *Mtb* to much higher concentrations than are seen in the clinic, creating considerable health risks for laboratory workers who are interacting with live micro-organisms. This work can only be undertaken in a technologically advanced environments such as Biosafety Level 3 laboratory which are expensive to build and maintain. LMICs are at obvious disadvantage in terms of access to such facilities [118].

2.6.3 Molecular microbiology tools

To overcome the obstacles of slow turnaround and high cost of *mycobacterial* culture, in recent years attention has shifted towards development of molecular tools which can rapidly detect genetic material, e.g. deoxyribonucleic acid (DNA), of *Mtb* cells directly in biological samples without the need to grow the organism. The most prominent example of this approach is the Xpert® MTB/RIF assay [24]. To use this technique in the diagnosis of pulmonary TB, sputum collected from presumptive TB patients is transferred into custom-designed cartridges. These are inserted to a machine which uses polymerase chain reaction (PCR) methods to detect and selectively amplify any *Mtb*-specific DNA present. Specific DNA sequences are used to report: i) whether any *Mtb* bacteria were present in the sputum, and ii) whether those bacteria were resistant or susceptible to rifampicin-based TB treatment. Over time, the Xpert® technology has been refined to improve its performance [32], and to report on drug resistance patterns to anti-TB drugs other than rifampicin [177].

The Xpert® molecular approach has several advantages over both smear microscopy and *mycobacterial* culture, as it is fully automated, produces results very quickly under optimal conditions, within 90 minutes of a sputum sample being inserted in the machine [24], and can perform DST in parallel with detection of bacteria. For these reasons, many clinicians and researchers now advocate replacing microscopy and culture with Xpert®-based tools for front-line TB diagnosis worldwide [6].

However, there are drawbacks to this approach too. Firstly, it is much more expensive than sputum smear microscopy; each cartridge for the machine costs $15 compared to the $0.10 cost of a microscopy slide for every sputum sample [6] which hampers roll-out in LMICs. The test is not available in many resource-poor locations. Secondly, small amounts of DNA from dead *Mtb* cells are detected in exactly the same way as viable bacteria, meaning that the assay stays positive long into the course of TB treatment, and cannot be used to monitor treatment response [78]; sputum smear microscopy remains the WHO-endorsed tool for this task [165].

Overall, whilst *mycobacterial* culture is currently the 'gold standard' TB diagnostic test it is too slow and expensive for front-line use in many locations. Xpert® MTB/RIF and other molecular tests are much faster and provide detailed read-outs including DST but are unavailable in some settings and are not recommended to monitor TB treatment response. Therefore, a role remains for sputum smear microscopy, but the accuracy of this method is subjectively dependent on operator-performance. Computer-assisted automation of microscopy including image analysis – particularly if tools could be developed which are implementable in LMICs – may extend and increase the value of this tool for TB diagnosis (as well as treatment monitoring and prognosis) and help accelerate overall progress towards achieving the End TB goals.

2.7 TB treatment

As outlined in Section 2.3, once a diagnosis of active TB disease is made, prompt initiation and successful completion of therapy is important to improve outcomes for individual patients (lowering mortality) and to assist in disease control by reducing transmission and preventing development of antibiotic resistance. This section will describe current approaches to TB treatment. As with Section 2.6, the specific focus will be on the management of pulmonary TB.

2.7.1 Challenges with current anti-TB treatment regimens

Standard WHO treatment for patients with TB is shown in Figure 2.8. Those with drug-susceptible disease (i.e., without any resistance to first-line antibiotics), require therapy for six months. A combination of four antibiotics (isoniazid, rifampicin, pyrazinamide, and ethambutol) is used for the two-month 'intensive' phase, followed by the four-month 'continuation phase' of two antibiotics (isoniazid, rifampicin) [165]. Patients with rifampicin-resistant TB, i.e. where the most important first-line antibiotic cannot be used have to take even longer treatment courses (up to 20 months) using five or more second line drugs [166].

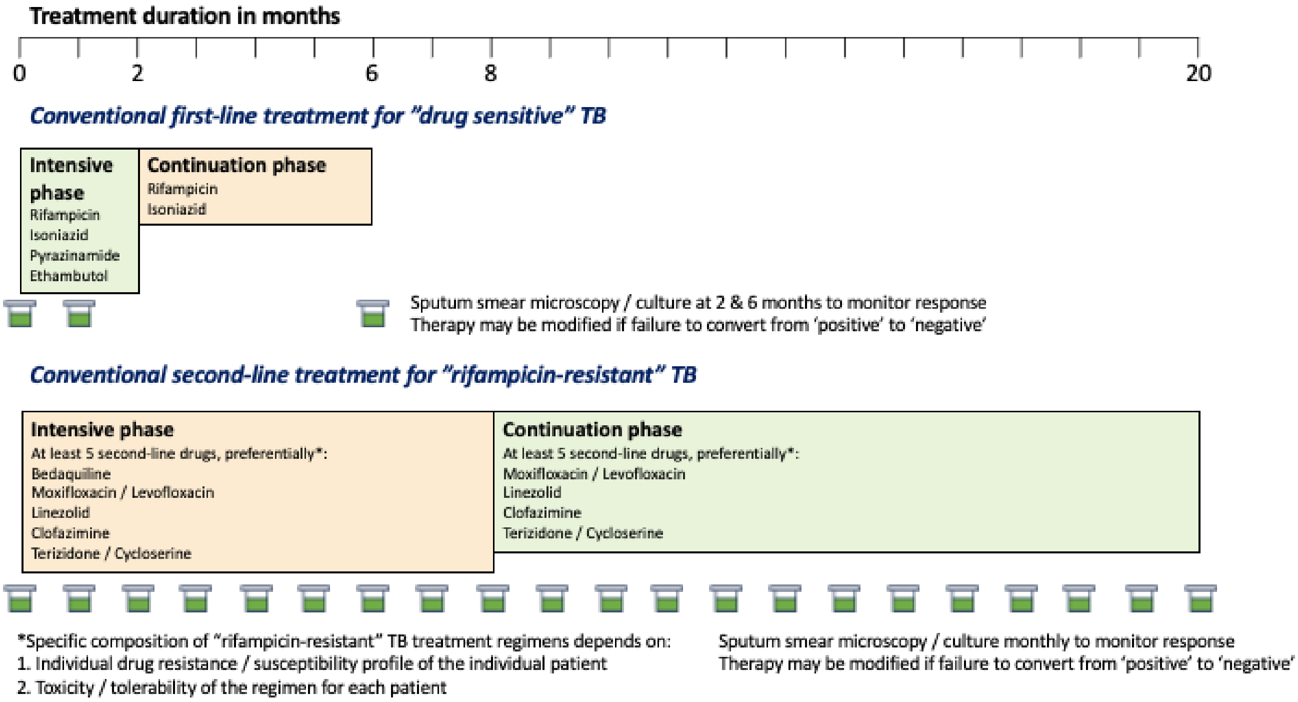

Treatment duration in months
0
2
6
8
20
Conventional first-line treatment for "drug sensitive" TB
Intensive phase
Rifampicin
Isoniazid
Pyrazinamide
Ethambutol
Continuation phase
Rifampicin
Isoniazid
Sputum smear microscopy / culture at 2 & 6 months to monitor response
Therapy may be modified if failure to convert from 'positive' to 'negative'
Conventional second-line treatment for "rifampicin-resistant" TB
Intensive phase
At least 5 second-line drugs, preferentially*:
Bedaquiline
Moxifloxacin / Levofloxacin
Linezolid
Clofazimine
Terizidone / Cycloserine
Continuation phase
At least 5 second-line drugs, preferentially*:
Moxifloxacin / Levofloxacin
Linezolid
Clofazimine
Terizidone / Cycloserine
*Specific composition of "rifampicin-resistant" TB treatment regimens depends on:
1. Individual drug resistance / susceptibility profile of the individual patient
2. Toxicity / tolerability of the regimen for each patient
Sputum smear microscopy / culture monthly to monitor response
Therapy may be modified if failure to convert from 'positive' to 'negative'

Figure 2.8: WHO recommended treatment regimens and monitoring for pulmonary TB

The mean reasons for using multi-drug combination therapy are to kill *Mtb* cells more effectively and prevent development of resistance from individual genetic mutations by targeting the bacteria simultaneously with agents that work via different mechanisms of action. For example, all of the 'first-line' drugs work in different but complementary ways, some targeting the bacterial cell wall and others targeting essential intracellular processes, including the metabolic pathways involved in synthesis of key proteins required for survival. This is shown in simplified form in Figure 2.9

Irrespective of the antibiotic combinations used, the long duration of TB treatment required is problematic for several reasons. Firstly, case-holding and supervision of therapy for 6-20 months requires a robust public health infrastructure which National TB Control Programmes in many high-burden LMICs do not have. Once patients start to feel better, they may decide to prematurely discontinue medication [13], particularly if there are financial or practical challenges to repeatedly visiting clinic. This increases the risk of relapse and development of drug resistance [205]. Secondly, multi-drug antibiotic combinations can cause dangerous side-effects. In first-line treatment for drug-susceptible TB: rifampicin, isoniazid, and pyrazinamide all potentially damage the liver [84], isoniazid may harm peripheral nerves [58], pyrazinamide is associated with joint pain [111], and ethambutol can cause visual problems [224]. Some of the second-line drugs used for rifampicin-resistant TB have even more complicated side-effect profiles [206]. Some

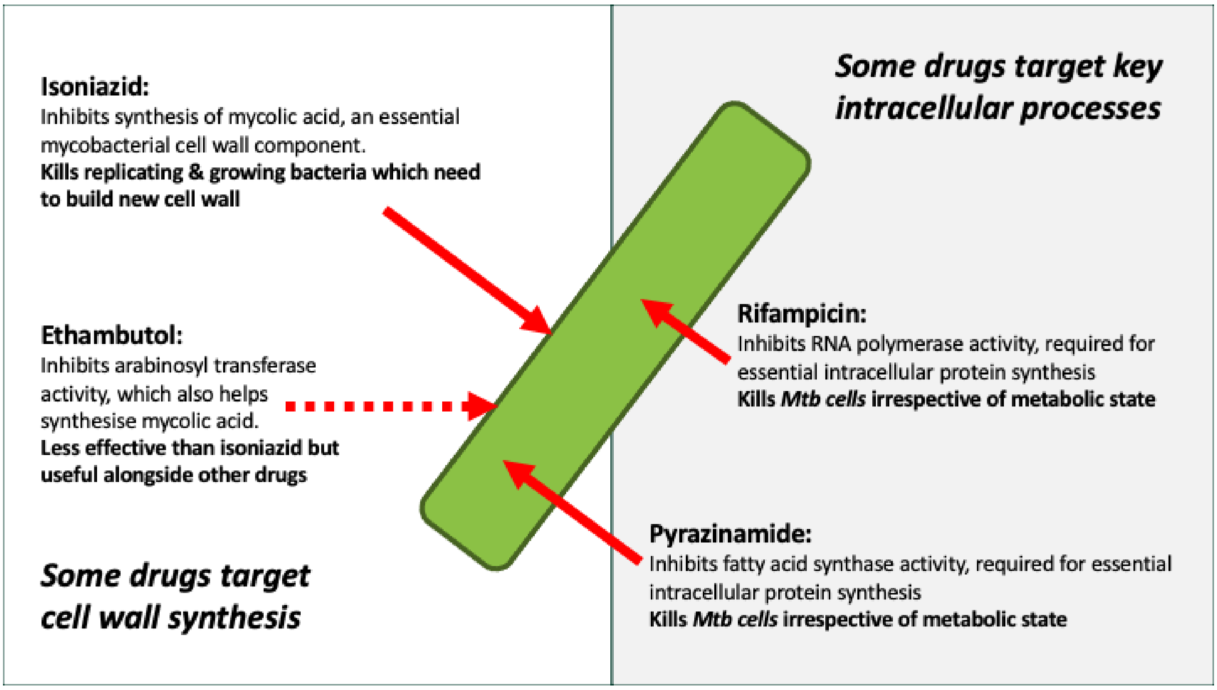

Isoniazid:
Inhibits synthesis of mycolic acid, an essential mycobacterial cell wall component.
Kills replicating & growing bacteria which need to build new cell wall
Ethambutol:
Inhibits arabinosyl transferase activity, which also helps synthesise mycolic acid.
Less effective than isoniazid but useful alongside other drugs
Some drugs target cell wall synthesis
Some drugs target key intracellular processes
Rifampicin:
Inhibits RNA polymerase activity, required for essential intracellular protein synthesis
Kills Mtb cells irrespective of metabolic state
Pyrazinamide:
Inhibits fatty acid synthase activity, required for essential intracellular protein synthesis
Kills Mtb cells irrespective of metabolic state

Figure 2.9: Main mechanisms of action of first-line anti-TB drugs

anti-TB drugs also interact with essential medications for other diseases (e.g. HIV) which may increase toxicity or reduce their effectiveness [156, 198]. All of these problems are progressively exacerbated if treatment has to go on for longer.

2.7.2 The need for careful treatment monitoring

Given that adherence to long-duration TB treatment is difficult, it is important for clinicians and patients to know that the antibiotics are effectively killing *Mtb* whilst they are taking it. Identifying individuals at high risk of treatment failure early allows investigation of the reasons for poor response and corrective action to be taken more quickly. At present, the microbiological tools available to monitor the effectiveness of pulmonary TB treatment are the same ones used for diagnosis: sputum smear microscopy, *mycobacterial* culture, and Xpert® MTB/RIF. Conversion from ‘positive’ to ‘negative’ on these tests, or falling bacterial burden, e.g. reduction from a smear ‘3+’ to ‘2+’ or ‘1+’ grading on Ziehl Neelsen or Auramine O microscopy, may be taken as evidence that treatment is working, and that long-term cure is likely if the course is finished. Alternatively, failure to convert to negative or rising bacterial burden may indicate than urgent changes to patient management are needed.

Both *mycobacterial* culture and Xpert® MTB/RIF have disadvantages as treatment monitoring tools. Although culture-based methods can precisely detect the presence or absence of live

Mtb cells in sputum, the lengthy time required to obtain results delays confirmation of effective treatment by several weeks post-sample collection. This wastes valuable time and may adversely affect outcomes. Whilst Xpert® MTB/RIF provides results quickly, degradation of *Mtb* DNA from dead bacteria is slow and the test can remain strongly positive for several months [78] creating unnecessary concern when treatment is progressing well. Although new treatment monitoring tools are under development [190], at present these problems have motivated ongoing WHO recommendations on the importance of sputum smear microscopy for monitoring of pulmonary TB therapy [165]. Therefore, computer-assisted automation of smear microscopy to improve the speed and objectivity of bacterial load estimation using this method may improve treatment monitoring as well as diagnosis.

2.7.3 Shortening TB treatment and the problem of non-replicating persistence

All patients suffering from TB have an undeniable need for therapy that is shorter and less hazardous. Recent clinical trials suggest that new regimens might reduce the duration of drug-susceptible and drug-resistant therapy to 4- and 6- months respectively [157, 42, 43, 65]. However, this will only partly relieve the problem. More research is urgently required to develop shorter TB treatments if the 'End TB Strategy' goals are to be reached.

Efforts to shorten TB therapy are more likely to be successful if they are based on a biological understanding of why current treatment takes so long. An important long-standing hypothesis in this area, illustrated in Figure 2.10, is that the *Mtb* organisms which cause TB disease in patients exist as a range of distinct sub-populations which, even when genetically identical, exhibit different phenotypic characteristics which cause differential responses to antibiotic exposure [213, 145].

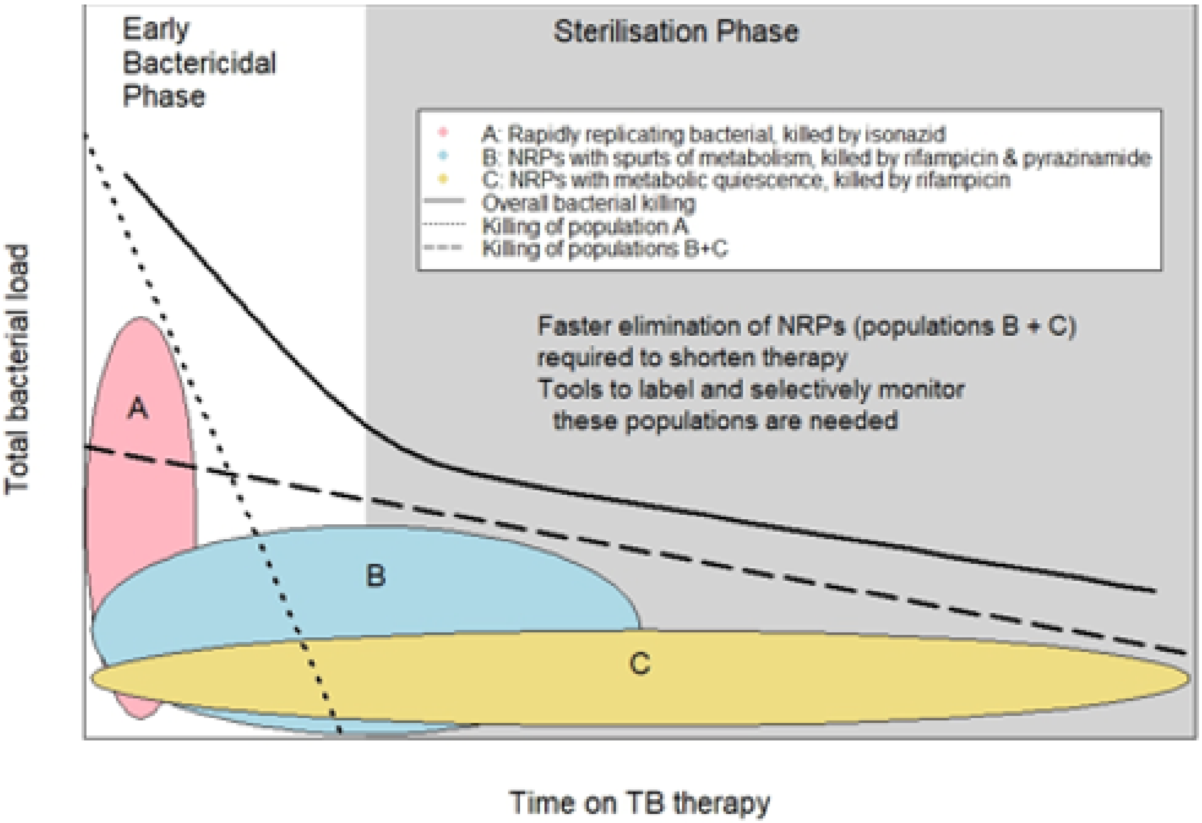


Figure 2.10: Bacterial killing of different populations of TB bacteria during treatment

Some metabolically active bacterial cells (Population A in Figure 2.10) replicate rapidly; these are easy to kill during an "early bactericidal phase" of treatment by antibiotics like isoniazid which disrupt their ability to synthesis new cell wall. However, other bacteria are quiescent and do not replicate at all for long periods of time. These cells, sometimes referred to as Non-Replicating Persisters (NRPs) are more antibiotic tolerant and harder to kill because they activate the metabolic pathways targeted by antibiotics less intensely (Populations B and C in Figure 2.10). Although some of drugs with intracellular mechanisms of action used in current anti-TB combinations such as rifampicin and pyrazinamide slowly kill NRPs during a "sterilisation phase" of treatment, a major priority in TB drug development is to develop medicines which selectively eliminate NRPs more quickly [272, 146]. Achieving this would be a major milestone in shortening TB treatment.

To help advance discovery and evaluation of drugs which kill NRPs, it is necessary to be able to distinguish drug tolerant persister *Mtb* cells from those that are easy to kill. *Mycobacterial* culture techniques cannot do this, because they can only measure the total quantity of viable bacilli. Existing molecular tests cannot do it either because drug-tolerance is a phenotypic rather than genotypic characteristic; the DNA-based sequences of metabolically active and quiescent cell populations is identical, but their behaviour is not. However, microscopy is able to look at individual bacteria one-by-one and there is some evidence that phenotypically distinct cell populations with different antibiotic response characteristics have different physical appearances which can be studied by direct visualisation. Two examples of this are outlined in the following sections.

2.7.3.1 Does bacterial length affect antibiotic response

Brief examination of sputum smear microscopy images, such as Figure 2.7, illustrate that *Mtb* cells vary in morphology: some are very short and straight whilst others are longer and sometimes curved. Although this variability is well described, the reasons for and implications of it have not been thoroughly investigated until recently [235]. Careful *in vitro* work using a microfluidic culture chamber to observe single cell growth and replication of *M smegmatis*–a non-pathogenic organism which is easy to work with, so is used to represent *Mtb* in some laboratory studies–has described three unusual and interesting characteristics of *mycobacterial growth* [39, 7].

Firstly, unlike other rod-shaped bacteria, the *mycobacterial* cell division cycle is governed by time and not size, i.e. each 'mother' cell divides into two 'daughter' cells after a specific time interval not when the mother cell reaches a specific length. Secondly, *mycobacteria* lack the molecular rulers which ensure symmetrical cell division by placing the division septum in the centre of the cell. Thirdly, *mycobacteria* lengthen asymmetrically from one pole, with faster elongation at older growth poles. As shown in Figure 2.11, this creates different types of cells at each division: one daughter cell inherits the already elongating pole and keeps growing on the same axis ("accelerator" cells), whilst the other daughter cell must generate a new growth pole and start to elongate more slowly on the opposite axis ("alternator" cells). By definition, in each successive generation of bacteria all alternator cells have new growth poles whilst accelerator cells inherit growth poles of varying ages with differing elongation rates.

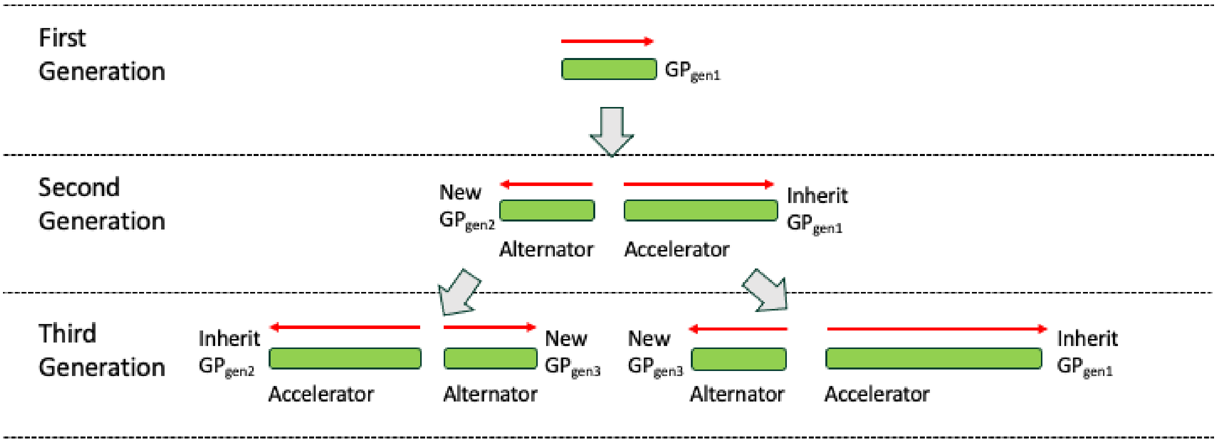


Figure 2.11: Arrows indicate the direction of unipolar growth. GP = growth pole, which preferentially occurs from the oldest extremity of the cell – the generation (Gen) of origin of each growth pole is show in subscript. Cells which continue with an existing GP to elongate on the same axis are termed “accelerators”. Cells which require to set up a new GP to elongate on the opposite axis are called “alternator”. Older GPs elongate faster than new ones.

These features go some way to explaining heterogeneity in the physical morphology of *mycobacterial* cells. Additional microfluidics work has also shown that bacteria of different sizes have differing antibiotic susceptibility. Longer birth length and mature growth poles are associated with rifampicin tolerance [184]. Accelerator cells appear proportionately more susceptible to killing by isoniazid than rifampicin, perhaps because more rapidly elongating cells need to build more cell wall, i.e. the process specifically targeted by isoniazid, whilst alternator cells are proportionately more susceptible to killing by rifampicin [7]. Translation of this *in vitro* work into investigation of variable cell length in *Mtb* from clinical sputum samples is just beginning [131]. Work on patients with multi-drug resistant TB from Vietnam showed, amongst other things, that increased cell length on Ziehl-Neelsen stained sputum smears was correlated with more severe disease [243]. A separate study, using a novel fluorescence microscopy method to analyse *Mtb* cells from the bloodstream of very ill HIV positive patients with advanced TB in South Africa, confirmed high inter-patient variability in bacterial cell length from clinical specimens and showed that mean bacterial length increased by 0.13 log-μm per day of TB treatment over the first three days [19].

Viewed together, all these findings suggest that cell length could be a useful marker of antibiotic tolerance and bacterial persistence. More research is needed to explore this possibility, but the work of taking detailed measurements from individual bacteria is extremely difficult and time-consuming, even on specimens from small patient cohorts. For example, the South African study described above was only able to analyse their bloodstream microscopy data on serial samples from 10 patients and the investigators from that project specifically noted that their method was "labour intensive and – as with all manual microscopy – requires subjective calls when classifying fluorescent objects as bacilli. . . Future goals should include automation and high throughout adaptatio" [19]. This observation extends the argument that computer-assisted automation of TB microscopy would be useful; not just for diagnosis and treatment monitoring but for research into deeper understanding of how different populations of *Mtb* cells respond to drug pressure.

2.7.3.2 Does bacterial lipid content affect antibiotic response?

In addition to differences in cell dimensions, it is likely that – if persistence is driven by variability in metabolic activity – phenotypic differences will be visible inside *Mtb* bacteria with differing degrees of antibiotic tolerance. For example, pathogenic *mycobacteria* often use fatty acids, which they derive from host cells, e.g. macrophages within the granuloma, as their energy source during human infection [152]. However, when exposed to physiological stress, e.g. low oxygen or excessively acidic environments, they can reduce energy consumption and divert these fatty acids into a biological pathway which deposits large lipid droplets in the cell cytoplasm [53, 14]. In simple terms, this may be a process similar to hibernation; in a hostile environment the organism 'shuts down' non-essential activities and store fuel for later.

Modification of the Auramine O microscopy method outlined in Section 2.6.1.2 allows discrimination between Lipid rich (LR) *mycobacterial* cells which are laden with triacylglycerol droplets and Lipid poor (LP) cells which are not [80, 207, 96]. In brief, after washing Auramine O from microscopy slides with acid-alcohol but before counter-staining with potassium permanganate, an additional fluorescence stain is used to label intra-cellular lipids. Typical stains for this purpose are Nile Red and LipidTOX Red neutral, both of which have longer wavelength excitation and emission spectra than Auramine O. When dual-stained slides are examined with a fluorescence microscopy, each FOV is viewed through a filter which identifies yellow/green Auramine O labelled *mycobacteria*, then through a separate filter which reports whether red lipid deposits are contained within them. Figure 2.12 shows discrimination between LR and LP *Mtb* cells within the same sputum smear using this technique.

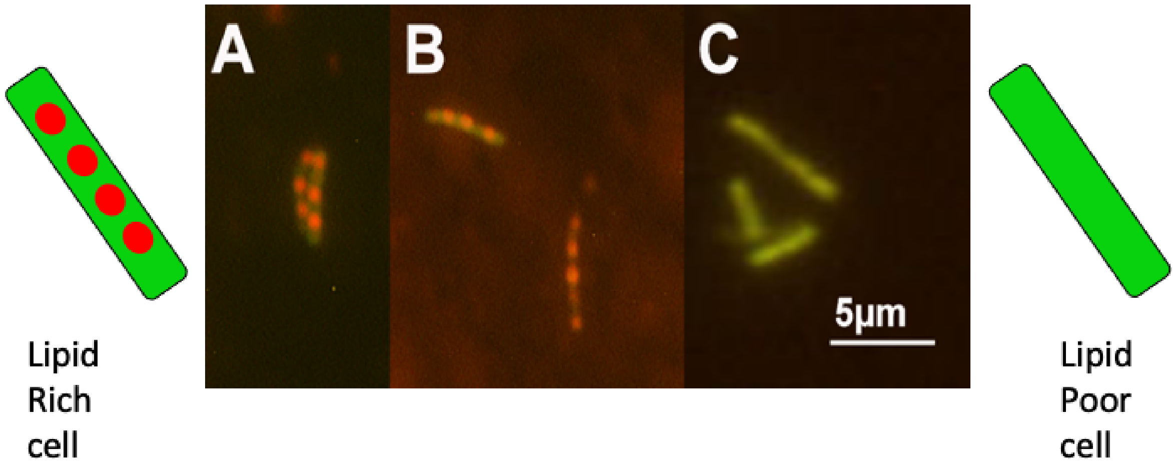


Figure 2.12: Images provided by Dr Sloan. Sputum smears from pulmonary TB patients stained using an Auramine O LipidTOX Red method and all images taken from ‘green’ and ‘red’ channel microscopy filters overlaid. In panels **A** & **B**, green Auramine O labelled *Mtb* cells are all LR, as evidenced by red fluorescence within them. In panel **C**, the *Mtb* cells are LP, as no red fluorescence is observed

There is considerable *in vitro* evidence that LR *Mtb* cells exhibit lower metabolic activity and reduced replication rates than LP cells [53, 80, 57]. As described above these features are generally believed to be beneficial for antibiotic tolerance. Additionally, the minimum inhibitory concentration of rifampicin and isoniazid which is required to kill *mycobacterial* cells has been shown to be higher for LR *mycobacteria*, confirming an association between cellular lipid content and antibiotic tolerance [96, 57].

Similar to the example of cell length, evidence of association between *Mtb* lipid content and treatment response from clinical studies is more limited. One study from Malawi reported that a median of 28% of *Mtb* cells in pre-treatment sputum samples from drug-susceptible pulmonary TB patients were LR. On serial sputum sampling over the first 28 days of therapy, the proportion of LR cells selectively increased over time amongst patients who went on to have unfavourable outcomes at 6 months [207]. This suggests that microscopically visible lipid content within *Mtb* is associated with antibiotic tolerance and linked to treatment failure, but the study was small; only 40 patients with serial microscopy data out to 28 days. This work has not yet been replicated, partly because manual assessment of the internal appearance of individual bacilli on images captured at multiple time-points is user-dependent and extremely laborious. Once again, this strengthens the argument for computer-assisted automation of microscopy to facilitate more robust investigation of relationships between phenotypically different *Mtb* populations and TB treatment response.

Chapter Three

Literature review on related work

Chapter abstract – This chapter reviews existing literature on AI methods to automate analysis of TB smear microscopy images. Variability in currently available image datasets and performance metrics used to evaluate key processes including image classification, regression and segmentation techniques are described. A comprehensive critique of all previous work using machine learning (ML) and DL approaches is presented, including an overview of strengths and limitations. Gaps in existing TB-AI microscopy methods research are highlighted and discussed.

3.1 Chapter introduction

Having outlined the rationale for automating analysis of TB microscopy images in Chapter 2, it is important to understand current work in the field before conceptualizing new approaches. This chapter will review the existing literature. Successful evaluation and deployment of any AI image analysis technique depends on the dataset used for the work. For TB microscopy the data are libraries of images derived from light / Ziehl-Neelsen (sometimes referred to as ‘brightfield’) or fluorescence (Auramine O) microscopy. Distinct image sets are often used to ‘train’ and ‘test’ new AI tools. Ideally, various methods ought to be consistently applied to new data and should be readily accessible for evaluation, accompanied by the standard performance metrics for each tool.

TB-AI for microscopy image analysis may be undertaken for a range of purposes; disease diagnosis, cell detection, bacterial load quantification (i.e. counting of cells once detected), or more detailed structural description (phenotyping) of individual bacteria. The effectiveness of methods designed for each purpose requires evaluation by appropriate metrics. It is important to understand what these metrics are and to consider how well they are standardised across existing methods so that different approaches to the same task can be compared.

Each TB-AI microscopy technique may have a foundation in different sub-fields of AI, often classical machine learning (ML) or DL. For the effective review and categorization of diverse methods in existing literature, comprehension of the foundational principles of conventional ML and DL, as well as their unique characteristics, is essential. ML is a subfield of AI focused on developing algorithms that allow computer systems to make decisions from data without explicit programming. ML enables computers to automatically discover patterns and insights from data rather than rely solely on rules-based instructions. It encompasses a range of methods including supervised learning (e.g. regression and classification) and unsupervised techniques (e.g. clustering, ensemble methods, and reinforcement learning.

DL is a subfield of ML that uses multi-layered artificial neural networks to learn data representations and patterns. The term "deep" refers to the multiple layers in these neural networks, which are composed of interconnected nodes, analogous to neurons in the human brain. DL models are trained to automatically construct complex concepts from simpler representations through transformations across hierarchies of layers. A core advantage of DL is the automatic learning of hierarchical feature representations directly from the raw data while preserving spatial information [122, 20]. This contrasts with conventional ML where method input may require considerable engineering to manually craft predictive features [94]. However, DL models demand large volumes of training data to unlock their representation learning potential [219]. In practice, data augmentation has reduced the extent of this problem. Generally, conventional ML remains effective for smaller datasets, but lacks the same feature extraction capacities. The dependency on data correlates with the complexity of the model; DL can address more complex tasks such as computer vision (CV) [113]. In contrast, conventional ML methods may be less suitable for these tasks, as they often lose spatial information. This loss occurs because the input usually takes the form of a vector, necessitating a prior flattening procedure. Overall, DL offers automated feature extraction and modeling of immense complexity, at the cost of substantial data requirements. Generally, conventional ML provides flexibility for small data, but with more constraints on feature engineering and model complexity.

CV is another subfield of AI focused on the computational interpretation of visual data. Contrasting with ML, which largely depends on pattern recognition from data, CV lacks a distinct learning phase. It employs mathematical algorithms that equip computers with the capability to "see" and execute tasks such as object recognition, image segmentation, and scene understanding. In practice, and in some published TB-AI microscopy methods, CV techniques are often combined with classical ML and DL methods as a hybrid strategy [159] to improve the quality of image representations and label generation. CV provides effective methods for tasks like low-level feature extraction. Meanwhile, conventional ML and DL models contribute pattern recognition and classification capabilities. Using both together can augment the strengths of each, leading to more robust performance than either alone. Combining ML/DL and CV techniques may provide a flexible framework for creating accurate representations to feed into downstream analytic tasks.

In this chapter I will review the existing literature on TB-AI microscopy, expanding on the above topics by addressing the following questions:

- How many datasets of TB microscopy images are available online, and what microscopy were used to generate them?

- What challenges to development of AI image analysis methods are presented by the level of variability in currently available TB microscopy image datasets?

- What performance metrics have previously been used to evaluate AI approaches to TB microscopy image analysis?

- What specific conventional ML and DL techniques, including those which are combined with CV, have been used for TB microscopy image analysis and how have they performed?

3.2 Methodology for literature review

To maximise the likelihood that all relevant material would be found, a standardised approach was taken to the literature search for this chapter. Detailed screening was undertaken of several academic databases. Systematic searches were undertaken in “PubMed”, “Scopus” and “Web of Science” using a combination of the following terms as keywords: “Tuberculosis” AND “Microscopy” AND “Automated (including Automation OR Pattern Recognition” OR “Image Processing” OR “Artificial Intelligence” OR “Deep Learning”. These searches were completed on 29$^{\text{th}}$ January 2023. The literature review initially incorporated more than the aforementioned keywords (i.e. precise technical terms such as ‘random tree forests’), leading to a subsequent narrowing of the search criteria to these specific terms to minimize the influx of irrelevant papers. For instance, including precise terms like CNN or random tree forests in the search parameters might result in retrieving a significant number of extraneous papers. These papers, while utilizing the mentioned techniques, focused on different medical tests, such as CXRs. Additionally, the search did not restrict the date range due to the relatively limited volume of literature in this field. Indeed, as will be discussed later, earlier works, such as that by Veropoulos *et al.* [242], have had a considerable influence on the research presented in subsequent chapters. Scientific publications identified from these searches were screened by title, abstract, and full text. Those which were written in English and described automated analysis of an original dataset of TB microscopy images were curated on a spreadsheet and duplicates were removed. The additional platforms "Google Scholar", "ResearchGate", "Academia", and "arXiv" were searched in a similar manner to identify relevant content within conference abstracts and grey literature documents which are not well represented in indexed databases. Reference lists and bibliographies from all papers selected for inclusion were also screened to identify any additional datasets which may have been missed by this search strategy. Some papers published in the course of experimental work described later in this thesis (authored by Zachariou *et al.*) were identified by my search but they have been excluded from this review, because its purpose is to describe work by other groups which have formed the background to my thinking.

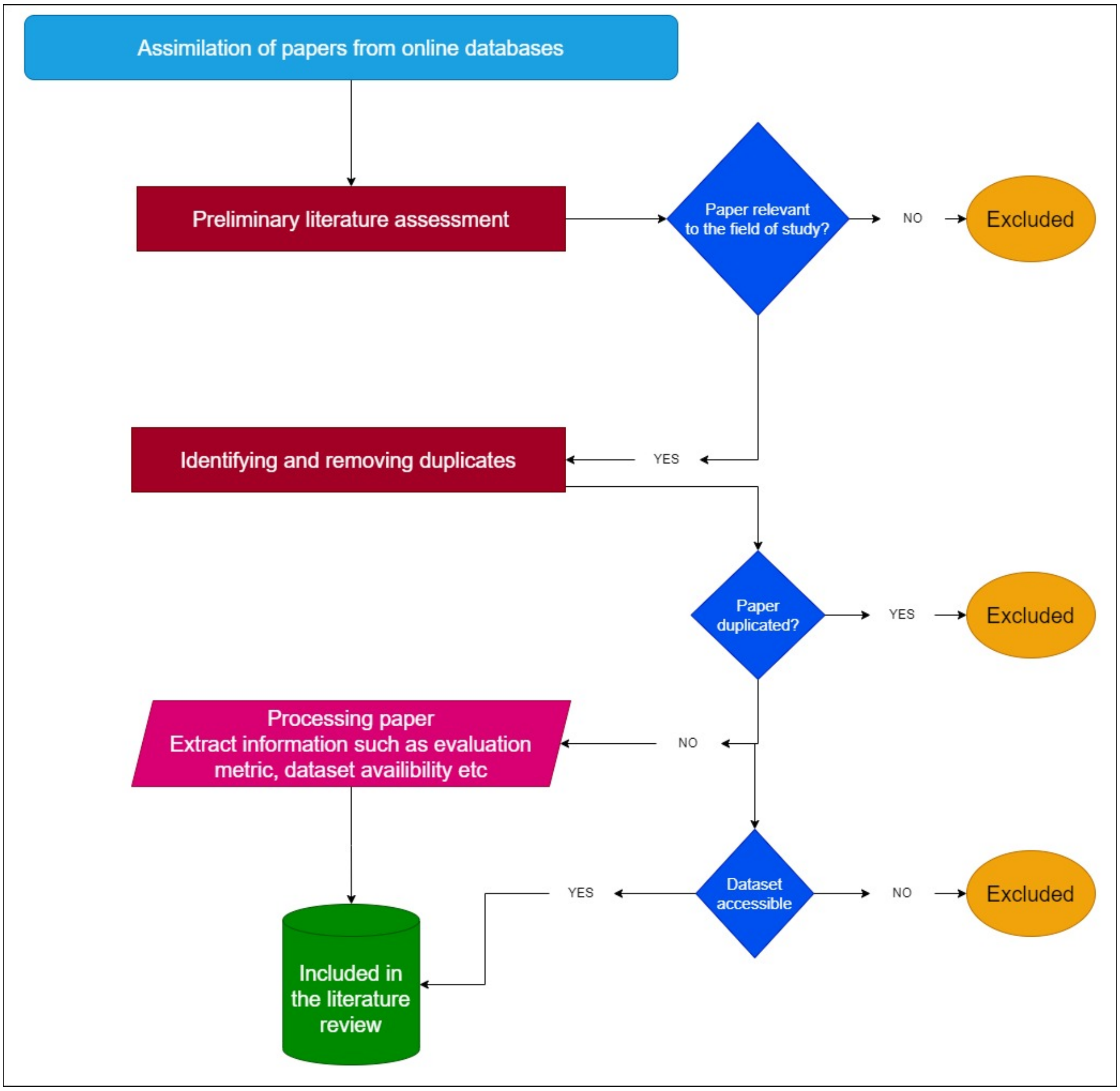

Assimilation of papers from online databases
Preliminary literature assessment
Paper relevant to the field of study?
NO
Excluded
YES
Identifying and removing duplicates
Paper duplicated?
YES
Excluded
NO
Processing paper
Extract information such as evaluation metric, dataset availibility etc
Dataset accessible
YES
NO
Excluded
Included in the literature review

Figure 3.1: Methodology used for literature search.

Once all relevant literature containing datasets were identified for inclusion, metadata were mapped for each, including whether the database used for analysis is openly accessible online. Supplementary metadata included: the specific microscopy technique employed (brightfield or fluorescence), the geographical origin of image generation and AI method development (according to United Nations geoscheme), the purpose of the research (diagnosis / detection of *Mtb* cells, quantification, or other), any quantitative performance metrics used, and whether the dataset for the work is publicly available online. Each published work was categorised according to whether the main AI approach was conventional ML or DL-based, with a simple discriminator being that all work which did not use convolutional neural network (CNN) or deep convolutional neural network (DCNN) as the core part of their method, were regarded as conventional ML whilst those that did employ CNN or DCNN were regarded as DL.

3.3 Description of available literature and datasets

Table 3.1 summarises original research work identified from the literature research.

As both light (brightfield) and fluorescence microscopy are used for *Mtb* bacterium visualisation, TB-AI research uses datasets of images derived from both methods. As described in Section 2.6.1, fluorescence microscopy generally offers better performance, but at higher costs. It has only become widely available in many centres over the last decade. A combination of these factors may explain why only 7/42 (17%) of all publications listed in Table 3.1 are based on fluorescence-based methods. Particularly in older methods, many image datasets are based on brightfield microscopy [137, 71]. 24/42 (57%) of methods used classical ML methods, whilst 18/42 (43%) used DL techniques, though DL approaches were more frequently used in works reporting after 2018. 36/42 (86%) of studied considered TB-AI microscopy methods as diagnostic tools or for *Mtb* detection. 6/42(14%) counted *Mtb* cells to quantify the sputum bacterial load, and none considered methods to phenotype individual bacterial cells. From my search, five databases of sputum smears for TB microscopy are currently accessible online. These are listed in Table 3.2 and are as follows: CDC Public Health Image Library (PHIL) [92], the Kaggle Tuberculosis image dataset [234], TB_IMAGES_DB_BACILLI V1 [47], Ziehl–Neelsen sputum smear microscopy image database (ZNSM–iDB) [200], and a dataset collected from brightfield microscopy sputum smears referred to by the authors as TBDB [233]. Only 5/42 (11%) methods listed in Table 3.1 used one of the openly available datasets described in Table 3.2. All other methods used of individually owned proprietary datasets. Of course, research based on proprietary datasets is valuable, but the lack of shared access to the raw image data creates challenges and reduces the transparency of comparative research between groups and methods.

Table 3.1: Summary of available literature on TB-AI microscopy.

Paper	Year	Microscopy Type	Region of image generation	Region of method development	Purpose of research	AI method used	Dataset online
Veropoulos *et al.* [242]	1998	Fluorescence	N/A	Europe	Diagnosis	ML	No
Forero-Vargas *et al.* [76]	2002	Brightfield	N/A	Europe	Detection	ML	No
Forero *et al.* [73]	2003	Fluorescence	Europe	Europe	Detection	ML	No
Forero *et al.* [75]	2004	Fluorescence	Europe	Europe	Detection	ML	No
Forero *et al.* [74]	2006	Fluorescence	Europe	Europe	Detection	ML	No
Sadaphal *et al.* [192]	2008	Brightfield	America	America	Detection	ML	Yes [92]
Costa *et al.* [48]	2008	Brightfield	America	America	Detection	ML	No
Makkapati *et al.* [137]	2009	Brightfield	N/A	Asia	Detection	ML	No
Sotaquŕa *et al.* [212]	2009	Brightfield	America	America	Quantification	DL	No
Khutalang *et al.* [115]	2009	Brightfield	Africa	Africa	Detection	ML	No
Khutalang *et al.* [116]	2010	Brightfield	Africa	Africa	Detection	ML	No
Osman *et al.* [169]	2010	Brightfield	Asia	Asia	Diagnosis	ML	No
Osman *et al.* [171]	2010	Brightfield	Asia	Asia	Diagnosis	ML	No
Osman *et al.* [170]	2010	Brightfield	Asia	Asia	Diagnosis	ML	No
Zhai *et al.* [270]	2010	Brightfield	N/A	Asia	Detection	ML	No
Nayak *et al.* [153]	2010	Brightfield	Asia	Asia	Quantification	DL	No
Chang *et al.* [34]	2012	Flueorescence	Africa	America	Diagnosis	ML	No
Costa-Filho *et al.* [49]	2012	Brightfield	America	America	Detection	ML	Yes [47]
Santiago-mozos *et al.* [197]	2014	Fluorescence	N/A	Europe	Diagnosis	ML	No
Ayas & Ekinci [11]	2014	Brightfield	Asia	Asia	Detection	ML	No
Costa-Filho *et al.* [50]	2015	Brightfield	America	America	Detection	ML	Yes [47]
Govindan *et al.* [91]	2015	Brightfield	America	Asia	Detection	ML	Yes (partially) [92]
Ayma & Castañeda [12]	2015	Brightfield	America	America	Detection	ML	No [92]
Gosh & Nasipuri [83]	2016	Brightfield	Asia	Asia	Diagnosis	ML	No
Priya *et al.* [179]	2016	Brightfield	Africa	Asia	Detection	ML	No
Soans *et al.* [209]	2016	Brightfield	N/A	Africa	Quantification	DL	No
López *et al.* [134]	2017	Brightfield	N/A	America	Detection	DL	No
Yan & Zhuang [255]	2018	Brightfield	Asia	Asia	Detection	ML	Yes [47]
Kant & Srivastava [114]	2018	Brightfield	N/A	Asia	Diagnosis	DL	No
Panicker *et al.* [174]	2018	Brightfield	America	Asia	Detection	DL	Yes
Samuel & Kanna [63]	2018	Brightfield	Asia	Asia	Detection	DL	Yes
Xiong *et al.* [254]	2018	Brightfield	Asia	Asia	Diagnosis	DL	No
Mithra & Emmanuel [147]	2018	Brightfield	Asia	Asia	Quantification	DL	Yes [200]
Díaz-Huerta *et al.* [60]	2019	Brightfield	America	America	Detection	ML	No
Ahmed *et al.* [4]	2019	Brightfield	N/A	Asia	Diagnosis	DL	No
Hu *et al.* [103]	2019	Brightfield	Asia	Asia	Diagnosis	DL	No
El-Melegy *et al.* [71]	2019	Brightfield	Asia	Africa	Detection	DL	No
Mithra & Emmanuel [147]	2019	Brightfield	Asia	Asia	Diagnosis	DL	Yes [200]
Vente *et al.* [241]	2019	Fluorescence	Africa	Europe	Quantification	DL	No
Yousefi *et al.* [257]	2020	Brightfield	N/A	America	Detection	ML	No
Serrão *et al.* [199]	2020	Brightfield	America	America	Detection	DL	No
Swetha *et al.* [221]	2020	Brightfield	N/A	Asia	Diagnosis	DL	No

Table 3.2: Details of currently accessible online sputum smear microscopy image datasets. The last column provides information about the manner in which the database represents various classes, if mentioned. Most annotated databases commonly utilise bounding boxes as a method for annotation. However, the TBDB database does not provide explicit documentation on how labels are constructed.

Image Dataset name	URL	Content of dataset	Image annotation	Label type
CDC Public Health Image Library [92]	`phil.cdc.gov`	Microscopy images within general collection of TB-related images, 25 brightfield slides 15 fluorescence slides	None	N/A
Kaggle Tuberculosis Image Dataset [234]	`kaggle.com/datasets/saife245/tuberculosis-image-datasets`	1265 brightfield images	Yes	Bounding Boxes
TB_IMAGES_DB_BACILLI.V1 [47]	Free access can be applied for at `tbimages.ufam.edu.br`	120 brighfield images	Yes	Bounding Boxes
ZNSM-iDB [200]	`drive.google.com/drive/folders/1HPcJzwKi76WwCFYj7dHUgVA31dAyFyTF`	9 sets of brightfield images (50-90 images per set)	Yes	Bounding Boxes
TBDB [233]	Freely available by contacting the authors	3102 brightfield images	Yes	Not specified

3.3.1 Challenges with dataset standardisation

Irrespective of the dataset used, a consistent challenge when analysing TB microscopy images is that the process of sputum smear preparation and image capture is hard to standardise. Figure 3.2 illustrates some key challenges.

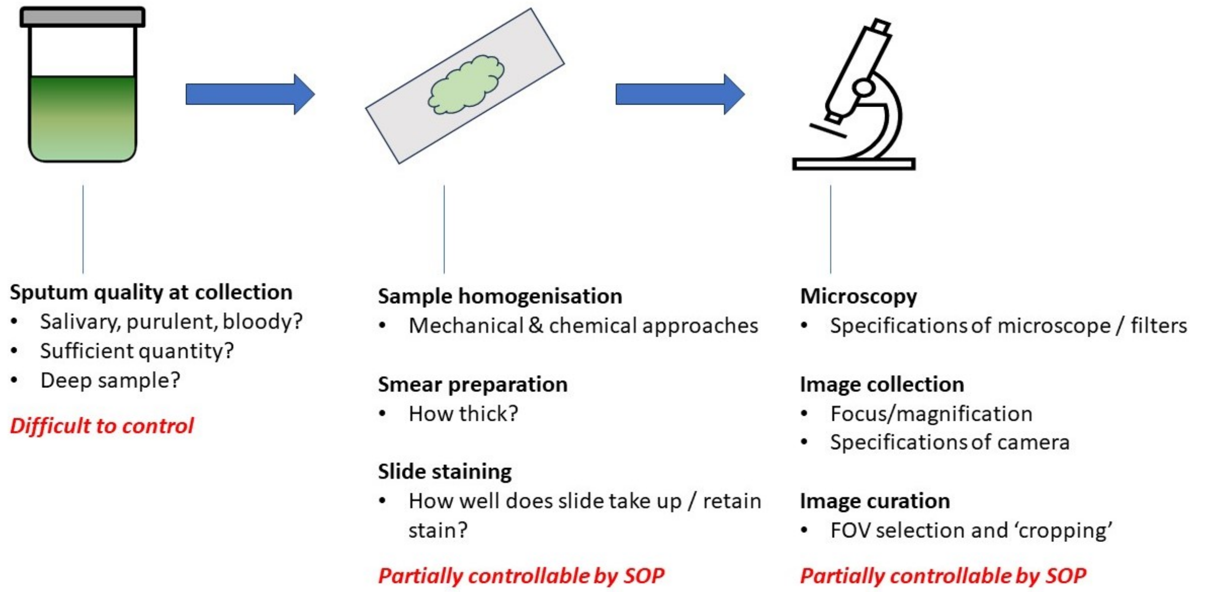

Sputum quality at collection
• Salivary, purulent, bloody?
• Sufficient quantity?
• Deep sample?
Difficult to control
Sample homogenisation
• Mechanical & chemical approaches
Smear preparation
• How thick?
Slide staining
• How well does slide take up / retain stain?
Partially controllable by SOP
Microscopy
• Specifications of microscope / filters
Image collection
• Focus/magnification
• Specifications of camera
Image curation
• FOV selection and 'cropping'
Partially controllable by SOP

Figure 3.2: Challenges to image dataset standardisation

Even when carefully written standard operating procedures (SOP) are meticulously followed, expectorated sputum is variable in consistency and difficult to homogenise. This affects the thickness of smears and influences the degree of background material and stain uptake on microscopy slides (see Figures 2.6 and 2.7). After slide preparation, the process of "reading" them comprises magnification (typically from ×400 to ×1000) and sequential examination of small FOVs. When researchers are preparing collections of FOVs for automated analysis, procedures vary. The most common options are i) manual inspection and creation of the image set [241], ii) auto-focus algorithms [75, 114, 270], or iii) successive cropping of the whole slide followed by a filtering stage to remove FOVs void of bacteria [103]. Individual FOVs, or sub-sections of them, are often additionally cropped into even smaller patches, where bacteria are present. All these methods ultimately utilise FOVs of arbitrary dimensions with no pre-specified standards for the width and height of each image. Furthermore, each image collection might differ in terms of spatial dot density, which alters the magnification levels of a bacterium's physical size. Researchers in different settings often have different hardware (e.g., different specifications of digital camera).

This all has implications for downstream biological research based on image interpretation. For example, Section 2.7.3 previously described the value of measuring the physical size of bacteria under different physiological or treatment conditions, but this is impossible if image dimension and magnification are not standardised and recorded. Studies using online accessible image sets illustrate this problem. Yan *et al.* [255] evaluated their approach to *Mtb* cell detection from Ziehl-Neelsen stained smears on both their own proprietary dataset and the online ZNSM–iDB dataset [200], with the latter yielding much lower accuracy because dimensions and resolution vary considerably within the ZNSM–iDB images.

Conventional ML, and DL, are data-driven, and most TB-AI microscopy methods studied to date suffer from insufficient training data, which substantially impacting performance. The fact that so few openly available TB smear microscopy datasets exist, with no standardisation in the methods used to generated them may have contributed to a situation where most publications use their own dataset both for training and testing of their methods. Such a constraint reduces the likelihood that a method can be seamlessly applied to a different dataset while achieving comparable results. Additionally, it renders the replication of results unfeasible if the original data is absent. Consideration should be given to whether it is possible to establish databases of microscopy images according to standardised protocols. Although desirable, this may be difficult to achieve because some of the causes of variability between datasets outlined above are hard to eliminate.

3.4 Evaluation of performance metrics

Assessing the performance of any new technique is done by performance metrics, so that novel approaches can be compared to the current state of the art or other experimental methods. In the literature on TB-AI microscopy, three general approaches are taken, dependent on the aim of the work: classification, regression, and segmentation. Common metrics to evaluate each of these are summarised in Table 3.3. These metrics are mainly used to convey empirical evidence of a model's performance on test data, but they also have utility during the learning phase on training data, serving as a form of cross-validation to verify the model's learning efficacy.

Table 3.3: Common performance metrics for evaluation of ML/DL tools in TB-AI microscopy

Approach to image analysis	Common application	Common evaluation metrics
Classification	Detecting of objects of interest of (Y/N) in a slide, FOV or patch	Accuracy Sensitivity (recall) Specificity Precision / Predictive Positive Rate F-measure
Regression	Counting objects of interest in a FOV	Mean average error Mean squared error R^2
Segmentation	Localising objects in the FOV. Sometimes precedes classification or regression.	Sorensen-Dice coefficient Jaccard index Hausdorff distance

3.4.1 Classification metrics

For many medical tests, the most important classification metric is the capacity to accurately differentiate between 'positive' and 'negative' occurrences. In TB microscopy this may be the ability to correctly diagnose a patient with pulmonary TB or the ability to correctly detect *Mtb* cells on a sputum smear. The approaches to diagnosis and detection are similar, the terms are often used interchangeably in the literature, and the same performance metrics are used for both. However, there are differences between these concepts which can influence assessment of whether a new method is working properly. Figure 3.3 illustrates this explanation in more detail. Overall, to assist with TB treatment monitoring, high performance metrics for detection of individual bacteria is desirable.

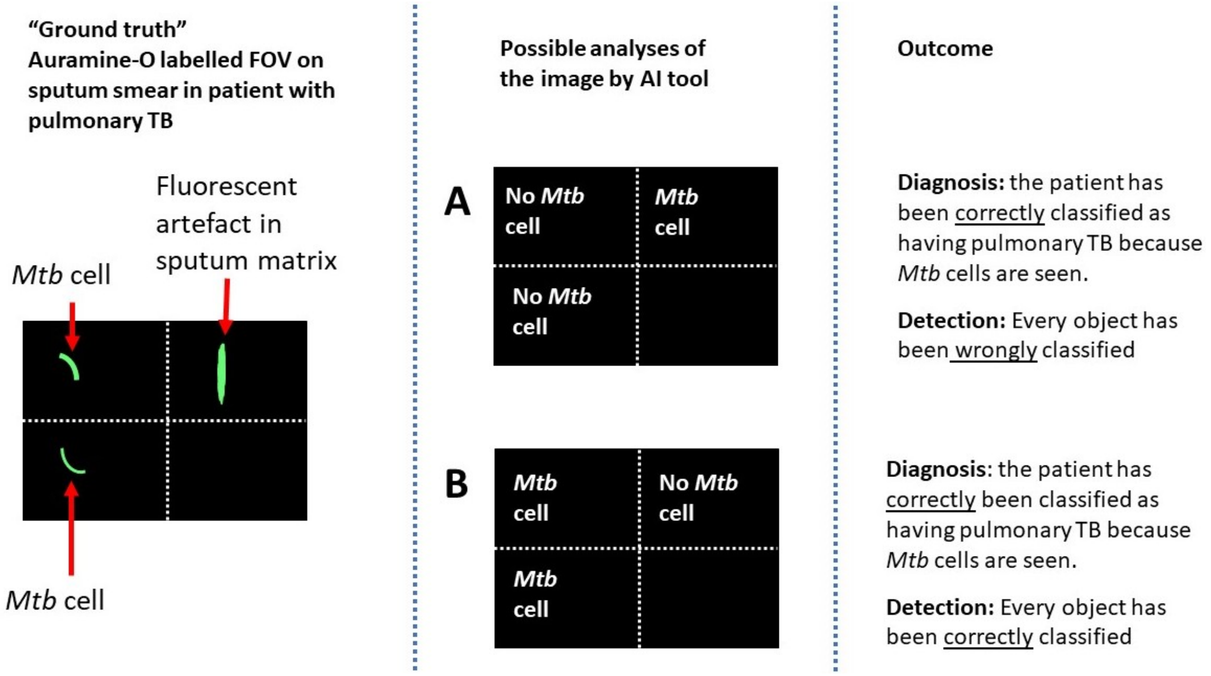

"Ground truth"
Auramine-O labelled FOV on sputum smear in patient with pulmonary TB
Fluorescent artefact in sputum matrix
Mtb cell
Mtb cell
Possible analyses of the image by AI tool
A
No Mtb cell
Mtb cell
No Mtb cell
B
Mtb cell
No Mtb cell
Mtb cell
Outcome
Diagnosis: the patient has been correctly classified as having pulmonary TB because Mtb cells are seen.
Detection: Every object has been wrongly classified
Diagnosis: the patient has correctly been classified as having pulmonary TB because Mtb cells are seen.
Detection: Every object has been correctly classified

Figure 3.3: In analysis A of this image, the correct diagnosis is reached because the sputum artefact is wrongly 'detected' as a bacterium, even though true bacteria are not detected. In analysis B all objects are correctly detected. Classification based only on 'diagnosis' here would over-report performance 'detection' for individual objects.

Performance metrics for classification procedures include accuracy, sensitivity (recall), specificity, and precision which all can be calculated from knowing the true positive (TP), true negative (TN), false positive (FP) and false negative (FN); their definition is shown in Table 3.4.

Table 3.4: Definitions of elements for performance metrics in model classification

Group	Description	Diagnosis / detection examples with reference to TB-AI microscopy
TP	An event / object of interest is correctly reported	A person with TB is correctly diagnosed with TB; fluorescence on an FOV is correctly identified as an *Mtb cell*
TN	The absence of an event / object is correctly reported	A person without TB has a negative test; a FOV containing no TB cells is reported as negative
FP	An event / object is reported in error	A person without TB is diagnosed with the disease; an artefact on a slide is incorrectly identified as an *Mtb* cell
FN	An event / object of interest is missed	A person with TB is not diagnosed; a FOV containing an *Mtb* cell is reported as containing background matrix only

Using these elements, the definition of accuracy is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.1}$$

Whilst accuracy can be useful, it has shortcomings when assessing diagnostic or detection models with imbalanced datasets, which is the situation for most medical AI applications. For example, most FOVs on TB microscopy slides contain background matrix, with *Mtb* bacteria scattered in a few locations. A model might report high accuracy solely based on its correct identification of TNs. However, such a model lacks utility if it consistently fails to accurately detect TPs.

Sensitivity (also known as recall or true positive rate (TPR)) and specificity are also frequently reported in TB-AI microscopy studies. Sensitivity assesses model performance by reporting the proportion of positive events over objects that are correctly classified and is defined as:

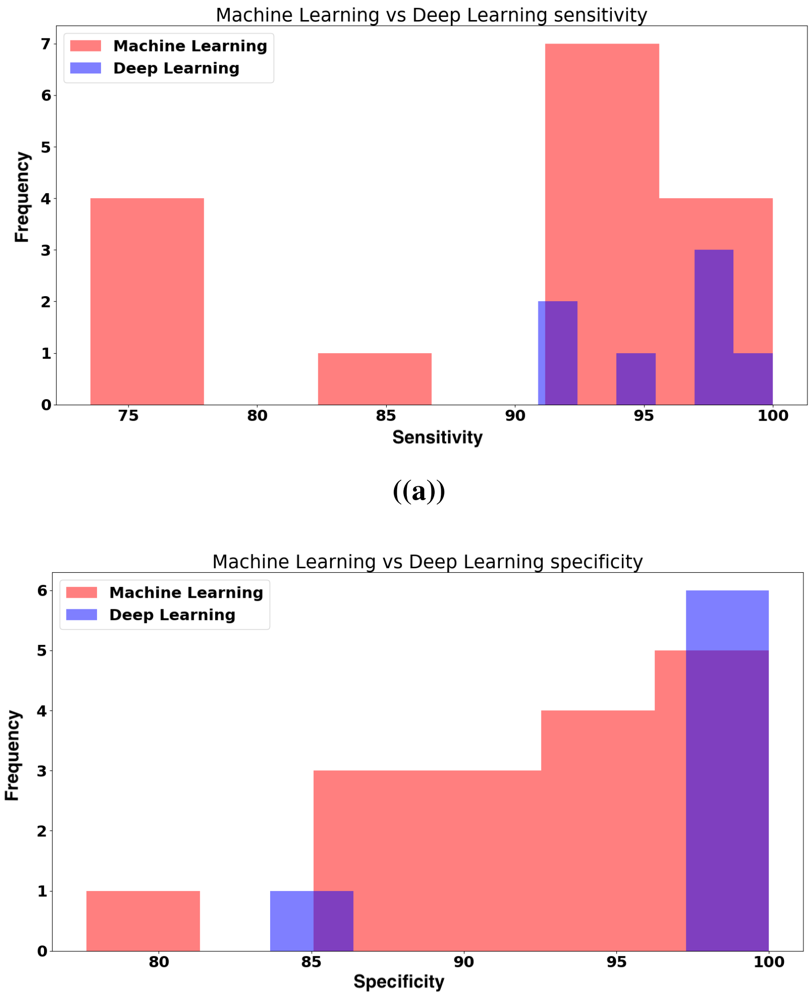
$$Sensitivity = \frac{TP}{TP+FN} \tag{3.2}$$

High sensitivity from TB sputum smear analysis could be interpreted to mean that very few *Mtb* cells were missed. Conversely, specificity (also known as true negative rate (TNR)) assesses model performance by describing the proportion of negative events over objects that are correctly reported and is defined as:

$$Specificity = \frac{TN}{TN+FP} \tag{3.3}$$

High specificity from TB sputum smear analysis could be interpreted to mean that very few artefacts in the sample were mis-reported as *Mtb* cells.

An interesting observation can be made by comparing conventional ML (n=16) and DL (n=7) methods from my literature search, and which used sensitivity and specificity metrics to report on classification models for TB diagnosis and/or *Mtb* detection. Figure 3.4. shows that sensitivity and specificity ranges attained by ML methods were variable (sensitivity: 75-100%, specificity: 80-100%) whilst equivalent ranges attained by DL methods were consistently higher (sensitivity: 90-100%, and specificity: 97-100%).



((a))

((b))

Figure 3.4: Comparative analysis of the sensitivity and specificity attained by works utilising ML and DL methods. Even though they are fewer in number, DL algorithms routinely score >90 on both sensitivity and specificity.

Precision (also sometimes referred to as the predictive positive rate (PPR)) measures the fraction of instances classified as positive that are actually true positives, and is defined as:

$$Precision = \frac{TP}{TP+FP} \tag{3.4}$$

The PPR can be informative when it is more important to correctly identify positive than negative objects or events [63, 134]. Interpretation of TB microscopy images arguably falls into this category because the ability of any approach to find a few positive FOVs amongst many negative ones on a sputum smear is fundamental to the usefulness of the method. The PPR is frequently used alongside receiver operating characteristics (ROC) and area under the ROC curve (AUC) metrics. These measurements are instrumental in analyzing the trade-off between the TPR (also known as sensitivity) and the FPR. This is particularly relevant for methods that produce continuous output variables and necessitate decision thresholds for binary classification. PPR is defined as:

$$FPR = \frac{FP}{FP+TN} \tag{3.5}$$

but can also be expressed as $1 - sensitivity$. ROC curves typically display TPR on the y-axis and FPR on the x-axis and the AUC is a scalar value which summarises the overall performance of the model at all classification thresholds. The AUC value ranges from 0 to 1, where:

AUC = 0.5 : The model, on average, fails to exhibit superior performance when compared to random guess.

AUC > 0.5 : The model, on average, outperforms random guessing, with greater AUC values indicating superior performance.

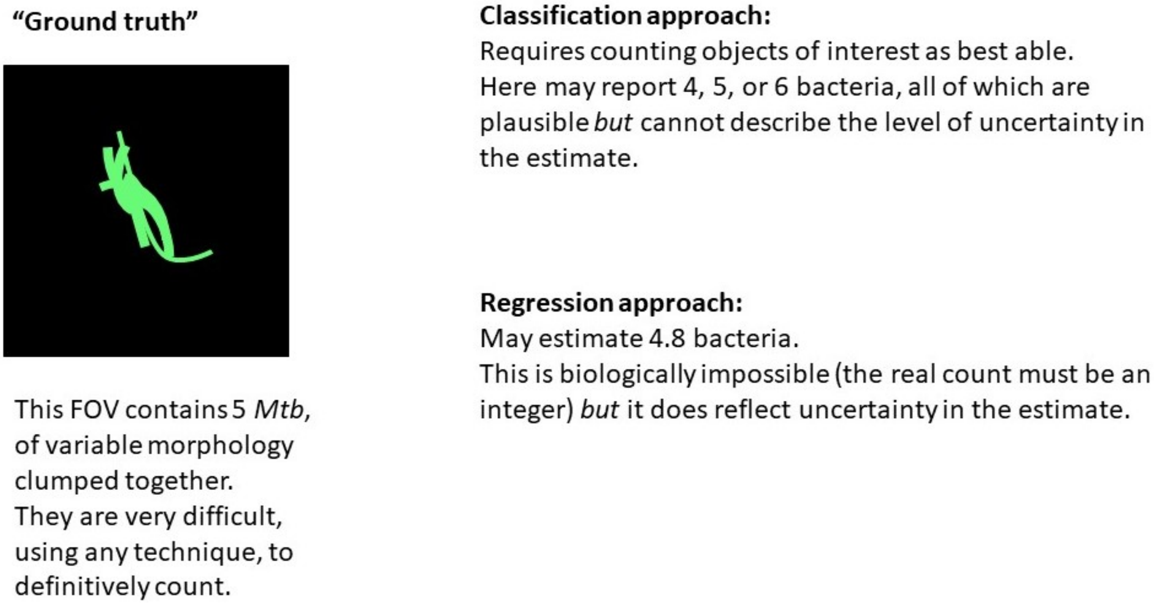
AUC = 1.0 : The model has perfect discriminatory power, achieving a true positive rate of 1 and a false positive rate of 0.

The F-measure (or F1-score) combines precision and sensitivity (recall) into a single metric that reflects the model's ability to accurately identify positive events or objects (precision) while also successfully capturing a high proportion of actual positive cases (recall). It is better for describing model performance than metrics such as accuracy in unbalanced datasets. To merge two metrics into once score, the F-measure is calculated as the harmonic mean of precision and recall and defined as:

$$F - measure = \frac{(2 * Precision * Recall)}{Precision + Recall} \tag{3.6}$$

3.4.2 Regression metrics

In Chapter 2, the value of counting bacteria in microscopic FOVs was explained: at diagnosis this information may reflect pulmonary TB severity and prognosis [108], whilst during treatment the rate of decline in bacterial load in serial sputum samples can help to track treatment response. Although object counting can be done by classification methods, regression analysis is often preferred as it is better able to preserve uncertainty in model estimates by predicting real numbers even when the actual count must be a natural number [241]. *Mtb* cells have a biological propensity to clump and counting them when they overlap is difficult so a modelling approach which describes some of the inherent uncertainty may be advantageous. Figure 3.5 illustrates this in a simple schematic and the issue will be discussed further in Chapter 5.



“Ground truth”

This FOV contains 5 *Mtb*, of variable morphology clumped together. They are very difficult, using any technique, to definitively count.

Classification approach:
Requires counting objects of interest as best able. Here may report 4, 5, or 6 bacteria, all of which are plausible *but* cannot describe the level of uncertainty in the estimate.

Regression approach:
May estimate 4.8 bacteria. This is biologically impossible (the real count must be an integer) *but* it does reflect uncertainty in the estimate.

Figure 3.5: Using regression to represent uncertainty in model-based estimates of *Mtb* counts

Typical metrics used to evaluate regression performance include mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2). Taking the average of all observations, MAE measures the absolute distance between the observations, e.g. ground truth bacterial load quantification in TB microscopy images, and regression predictions for the same measurement. The absolute value of the distances is used to correctly account for negative errors. MAE is expressed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{y}_i^{true} - \mathrm{y}_i^{pred}\right| \tag{3.7}$$

The mean absolute percentage error (MAPE) is a metric that is used to assess the accuracy of predictions by calculating the absolute percentage error for each data point. This error is determined by taking the absolute difference between the true value and the predicted value, and then dividing it by the true value. Subsequently, the formula computes the mean absolute percentage errors for each individual data point, followed by the multiplication of this average by a factor of 100 in order to represent the error in a percentage format. The MAPE is an essential metric for assessing the performance of prediction models, particularly in situations where datasets are challenging to normalise and have a broad spectrum of numerical values. However, MAPE can be highly sensitive to outliers, as it assigns equal weight to all errors, regardless of their magnitude. When extreme values are present, these can disproportionately influence the MAPE, leading to potentially misleading interpretations of a model’s performance [90]. The

MAPE is defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\mathrm{y}_i^{true} - \mathrm{y}_i^{pred}}{\mathrm{y}_i^{true}}\right| \times 100 \tag{3.8}$$

In the case of the MSE, the differences between observations and regression predictions are squared. This squaring ensures differentiability across all outcomes, making it well-suited for optimization techniques. Unlike the MAE, which is not differentiable at zero due to its use of absolute distances, the MSE allows for the calculation of gradients. Gradients are crucial in many optimization algorithms, as they guide the search for optimal parameter values. In the MSE, the absolute differences are replaced with squared differences, providing a smooth and differentiable loss function. MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{y}_i^{true} - \mathrm{y}_i^{pred}\right)^2 \tag{3.9}$$

The key difference between MSE and MAE is how each penalises mistakes when comparing predicted data to ground truth data. Since the MSE is a squared error, it penalises large errors more heavily. Consequently, MSE is more sensitive than MAE to outliers. The robustness of each metric and when it should be used is contingent on the nature of the task.

The root-mean square error (RMSE) is derived from the MSE in a manner analogous to the relationship between MAE and MAPE. The inclusion of the square root operation in the RMSE calculation guarantees that the resulting value is expressed in the same units as the original data. This enhances the interpretability of the RMSE and facilitates its comparison to the scale of the target variable. RMSE is expressed as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{y}_i^{true} - \mathrm{y}_i^{pred})^2} \tag{3.10}$$

Finally, R^2 represents the fraction of the variance in the dependent variable that a linear regression model explains. It is a scale-free score, so irrespective of whether the numbers are low or high, R^2 will always range from 0-1. It indicates the predictor variables' ability to explain variation in the response variable, i.e. how well the independent variables explain the dependent variable. Values closer to one indicate higher predictive ability. R^2 can be expressed as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{3.11}$$

R^2 always increases as more independent predictor variables are added, which may lead to the inclusion of redundant variables in the regression model.

3.4.3 Segmentation metrics

Segmentation refers to processes which are sometimes applied to FOVs in TB-AI microscopy to localise smaller regions of interest (patches containing possible *Mtb* bacteria) prior to classification or regression. The classification or regression models are then left with less complexity to manage, because they are applied to small patches rather than the whole FOV. Similar to classification tasks, pixel-wise accuracy assesses the extent to which each pixel is accurately assigned to its correct category. While pixel-wise accuracy is a commonly used metric for evaluating segmentation tasks, it may not be the most suitable measure in the context of microscopy FOVs, which often contain a majority of background pixels. A model could achieve a high pixel-wise accuracy by simply predicting most pixels as background, even if it fails to detect *Mtb* cells. Therefore, the model may become biased toward detecting the background rather than the cells of interest. Alternative metrics, such as the Jaccard index (J) or the Sorensen-dice (SD) coefficient, may offer a more balanced evaluation of the model's performance in segmentation tasks.

Explaining these metrics, requires a brief outline of the terms 'intersection' and 'union', including how they may be interpreted in the context of TB-AI microscopy. If S_1 and S_2 are the number of elements in a dataset, i.e. for digital TB-AI images, the pixel values in ground-truth and model-predicted segments of a FOV, the intersection of the two images ($S_1 \cap S_2$) is the set of common elements, i.e. overlapping pixel area, across the two segments, whilst union ($S_1 \cup S_2$) is the set of all elements in both segments (i.e., total combined pixel area of S_1 and S_2). The SD coefficient is defined as:

$$SD = \frac{2 \times |S_1 \cap S_2|}{|S_1| + |S_2|} \tag{3.12}$$

and represents the number of elements in the intersection of both S_1 and S_2 divided by the total number of elements in S_1 and S_2 combined. The Jaccard index is expressed directly as the ratio of intersection over union:

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{3.13}$$

For both SD and J, higher values (range: 0-1) indicate greater overlap between ground truth and model-predicted segments. These metrics are related. Given a value for SD, it is possible to determine the corresponding value of J, and vice versa. Nonetheless, these metrics differ and fulfill distinct evaluation objectives in the context of image segmentation. In general, J metric punishes single occurrences of incorrect classification more than the SD, even when both metrics agree that a single case is incorrect.

Evaluation of the proximity of a bacillus' perimeter is an alternative performance metric for

Table 3.5: Table displaying several assessment measures used by each publication. Each metric is separated by commas inside the metric column, and its associated quantity is listed in the value column. To be included in this table, publications must i) conduct a segmentation step and ii) give an assessment measure with an official value.

Paper	Segmentation Evaluation Metrics		
	Hausdorff distance	Jaccard Index	SD
Khutlang *et al.* [115]	0.96	N/A	N/A
Soans *et al.* [209]	0.06	N/A	87%
Diaz-Huerta *et al.* [60]	N/A	96%	N/A
Mithra & Sam Emmanuel [148]	N/A	95%	N/A

segmentation. The Hausdorff distance is a measure that calculates the distance between two subsets within a metric space. Given model predicted images with highlighted objects of interest and matching ground truth images, the closer distances between these boundaries reflect closer similarity. It does so by converting the set of non-empty compact subsets within a metric space into a metric space of its own. Specifically, when applied to two sets of points, denoted as S_1 and S_2, the Hausdorff distance is defined as follows:

$$H(S_1,S_2) = max(h(S_1,S_2),h(S_2,S_1)) \tag{3.14}$$

where $h(S_1,S_2)$:

$$h(S_1,S_2) = \underset{s_1 \in S_1}{max} \underset{s_2 \in S_2}{min} \|s_1 - s_2\| \tag{3.15}$$

is the directed Hausdorff distance between S_1 to S_2. The metric requires some underlying norm to be defined ($\|\cdot\|$); the L2 (or Euclidean distance) is typically employed. In some cases, the traditional Hausdorff distance may lead to skewed performance evaluations because it is sensitive to individual outliers. Furthermore, various methods have employed the Hausdorff distance for the purpose of image comparison, notably within the scope of the Modified William index (MWI) [33]. The MWI is a similarity index that combines the Hausdorff distance and the mean absolute distance between two sets of points or regions in an image. The authors used the Hausdorff in their work to determine the distance between each predicted structure and the actual structure in a given set of images [115]. Table 3.5 offers a summary of all the works along with their choice of segmentation metrics and results.

3.5 TB-AI microscopy research utilising ML

This section will provide a synopsis of all the research identified from my literature search which employed conventional ML algorithms in conjunction with CV techniques for classification and segmentation of TB-AI microscopy datasets. For simplicity, anything that does not adhere to the framework of deep convolutional neural networks (DCNN) is regarded as classical ML and included in this section. Approaches are thematically categorised into those which are 'local feature extraction' and those which are 'extracted pixel distribution'. Table 3.6 shows a summary of the most common evaluation metrics used by all papers included in this section. The degree of variability in the metrics chosen exemplifies the absence of an established standard for evaluation of new methods and underlines the difficulty in comparing research carried out in different settings.

3.5.1 Local feature extraction approaches

The underpinning local feature extraction ML approaches is to employ an edge or ridge detector to extract gradient intensities in the spatial domain of an image or a colour space threshold by determining bacteria pixel values beforehand. The color space threshold approach employs histograms to perform quantitative analysis on pixel hue bands, thereby identifying the typical color range where *Mtb* cells are most likely to be found. This is often the initial stage of a procedure which eventually aims to convert images into binary masks, fully erasing the background and leaving likely *Mtb* as white contours. Thereafter, a shape descriptor is employed to extract characteristics about the shape of probable bacilli. Subsequently, some researchers opt to use a classifier while others use heuristic information to make final decisions on whether the shape identified is *Mtb* or not.

From my literature search, Veropoulos *et al.* [242] published the first advance towards automated TB diagnosis using fluorescence microscopy images. They devised a five-step methodology, combining CV techniques with a simple neural network as a classifier. First, a Canny edge detector was employed to detect edges which enhances the images and extracts low-level features. Pixel linking served as a corrective measure to repair structural distortions induced by noise. Subsequently, the processed image underwent a transformation from the spatial domain to the frequency domain using the discrete Fourier transform (DFT). Fourier coefficients were calculated to serve as shape descriptors for bacteria and these were input into four kinds of classifiers: $k-$nearest neighbours, a neural network, kernel-adatron [77], and support vector machine (SVM). The model that performed most effectively achieved an accuracy of 97.9%. Even though this work considered bacteria very simply (i.e. singular elongated structures), it demonstrated the feasibility of *Mtb* detection using computer-aided image analysis.

Table 3.6: Performance metrics used to evaluate ML methods in TB-AI microscopy

Paper	Accuracy	Sensitivity/Recall	Specificity
Veropoulos *et al.* [242]	97.90%	94.10%	99.10%
Forero-Vargas *et al.* [76]	N/A	N/A	91.00%
Forero *et al.* [73]	N/A	93.30%	91.68%
Forero *et al.* [75]	N/A	86.66%	99.74%
Forero *et al.* [74]	N/A	94.67%	98.10%
Sadaphal *et al.* [192]	N/A	N/A	N/A
Costa *et al.* [48]	N/A	76.65%	88.65%
Makkapati *et al.* [137]	N/A	N/A	N/A
Khutalang *et al.* [115]	86.85%	99.95%	77.62%
Khutalang *et al.* [116]	93.47%	90.88%	95.85%
Osman *et al.* [171]	86.32%	N/A	N/A
Osman *et al.* [169]	98.07%	100.00%	96.19%
Osman *et al.* [170]	N/A	N/A	N/A
Zhai *et al.* [270]	N/A	100.00%	94.00%
Chang *et al.* [34]	N/A	92.30%	88.00%
Santiago-Mozos *et al.* [197]	N/A	73.53%	99.99%
Ayas *et al.* [11]	N/A	75.77%	96.97%
Costa-Filho *et al.* [51]	91.45%	93.41%	89.50%
Costa-Filho *et al.* [50]	93.25%	93.75%	88.46%
Govindan *et al.* [91]	N/A	72.89%	N/A
Ghosh *et al.* [83]	N/A	93.90%	88.20%
Priya *et al.* [179]	91.30%	91.59%	88.46%
Aymas & Castañeda [12]	70.52%	N/A	N/A
Yan *et al.* [255]	N/A	97.46%	93.99%
Díaz-Huerta *et al.* [60]	98.66%	N/A	N/A
Yousefi *et al.* [257]	82.27%	75.99	92.58

Santiago-Mozos *et al.* also used the Canny edge detector [112] as their primary approach to identify bacteria from fluorescence images but included an extra pre-processing step with an adjustable colour threshold for the green colour component of the image [197]. They used two successive SVM classifiers: the first discarded incorrectly identified objects from the previous stage and the second classified these objects based on their appearance. The first classifier employed a collection of rotation and translation-invariant characteristics of each candidate object as input. Forero *et al.* used a similar method, which included a segmentation phase comprised of a Canny edge detector, morphological operators, and classification of the resulting image [73]. Different bacilli characterisation and the use of clustering approaches for classification were major variations from earlier approaches. Follow-up work by the same group used a similar strategy, i.e. low-level feature extraction as input to a classifier, but mathematical autofocus algorithms were also utilised for image magnification and construction of FOVs [75]. Although the results either equalled or were slightly inferior to those from their previous work [73] and Veropoulos' work [242], this research was pioneering in implementing automated FOV generation, a topic pertinent to Section 3.8 of this chapter. Next, this group published work which performed low-level feature extraction through adaptive color thresholding and subsequently classification through clustering algorithms [74]. This time they used Gaussian mixture models since they were able to create the distribution of class features. FOVs for this dataset were prepared manually.

Using Ziehl-Neelsen stained images viewed by light microscopy, Costa *et al.* used a colour threshold-based segmentation stage, in which the authors were able to isolate bacteria from an image background [48]. For their proposed bacilli segmentation method they employed the use of Red minus Green (R-G) images from red-green-blue (RGB) format images and determined a threshold value that distinguished objects of interest from the background. Classification is conducted manually in this case, relying on heuristic features specifically crafted by the authors rather than employing ML classifiers. Makkapati *et al.* also used colour thresholding of light microscopy images as their principal segmentation technique [137]. Their proposed approach was to select the hue range $x^0 - 360^\circ$ on the HSV color space, where x is an adaptable number dependent on the input image. No classification technique was used, only a filtering stage utilising heuristic knowledge of bacterial morphology. The authors reported no performance metrics. In both of these works, the absence of an automated classifier had a harmful impact on the results in comparison to earlier methods [242, 73, 75, 74]. Two approaches for segmenting TB images using chromatic information were shown by Forero-Vargas *et al.* in work that did not include a classifier [76]. The first technique was based on the information contained in each distinct chromatic histogram and the fuzzy segmentation of colour images. The second was a straightforward colour filtering comparison of the inverse of the yellow-stained bacteria (blue channel) with the product of the other two chromatic channels.

Osman and colleagues published three similar methods [169, 171, 170]. The approach employs $k-$means clustering on the green component of the RGB color model and the R_y component of the C-Y color model to isolate tuberculosis bacilli from the background, which remains red despite the decolorization process [171]. Subsequently, a 5×5 median filter and region growing techniques were applied to remove small regions and noise. Although no evaluation metrics were provided, visual results indicated that some background still surrounded the bacteria in images. Secondly, using their segmentation technique to expand on their prior paper, the resulting segmented image was clustered into background and non-background regions [169]. After calculating the moments of the second and third order, a set of seven Hu invariant moments was extracted. These features were fed into a genetic algorithm neural network (GA-NN) for classification. Here, the authors reported 88.54% accuracy, but no sensitivity or specificity, for correctly classifying bacteria. In their third article, they used the same segmentation method as in the first, but this time employed the geometrical characteristics of Zernike moments [170]. Additionally, a hybrid multi layered perceptron (HMLP) was used for their final classification stage, which includes an extra connection to a layer beyond the immediately subsequent one. For the segmentation stage, the algorithm employed a dual-stage technique that utilizes both the HSV and CIE $L*a*b*$ color spaces, the latter having an adaptive threshold on the L component [270]. To classify the tuberculosis bacilli, three shape descriptors – area, compactness, and roughness – were used as feature extraction input to a decision tree for classification.

Remaining within the scope of colour space transformation, Costa-Filho *et al.* took a three-step approach [49]. Initially, they created a scalar selection from the following colour spaces: RGB, HSI, YCbCr, and Lab. The components and removal of components of these colour spaces were employed for pixel classification in the segmentation step. Then, a feedforward neural network pixel classifier with selected features as inputs was used to separate bacilli pixels from background. In the third stage, geometric properties, particularly eccentricity, and a newly proposed colour-based property, colour ratio, were employed for noise filtering. Using their technique from the first step, these authors released a second paper with the addition of three filters that use RGB space components: rule-based, geometric, and size filters [50]. The combination was then utilised as an input for an SVM and Neural network (NN). In this work, the authors improved their sensitivity results from 91.5% to 96.80%. Yan *et al.* retrieved a channel from Lab space and then extracted the edges (bacterial structures) using a gradient threshold [255]. In addition, aspect ratio, circularity, and area were employed to eliminate incorrectly detected structures.

Using just the RGB space, they defined conditions on each different component of the space that best meet the criteria for distinguishing bacteria from the background in a binary image [83].

To eliminate false contours (i.e. candidate regions of interest that were not bacteria) the shape, colour, and granularity features of the predicted contours were computed. Consequently, they used a fuzzy classifier in conjunction with the previously calculated characteristics to determine if a particular contour belongs to the class of bacteria or not. Priya *et al.* employed an active contour technique for their segmentation, which may be described as the application of energy forces and restrictions to separate the pixels of interest for further processing and analysis [179]. Following the segmentation of the image, the contours of the regions of interest were described using 15 Fourier descriptors (FD). The most significant of these descriptors were identified through the application of fuzzy entropy metrics. These FD were input to the SVM learning algorithm of a support vector neural network (SVNN)).

Yousefi *et al.* proposed a statistical model to detect *Mtb* bacilli based on their form and colour in Ziehl-Neelsen stained light microscope images [257]. These basic statistical models were used as a universal library for rebuilding any bacillus with different background colours and could overcome the challenges associated with geometric feature extraction techniques. Based on the extracted eigenvalues from the statistical models matrices, the authors used several approaches to classify individual bacilli and overlapping bacilli from the rest of the image. The $k-$NN classifier performed best amongst the evaluated classifiers, with an average accuracy of 82.7% for overlapping bacilli detection. In addition, the accuracy of their method for detecting single occurrences of bacteria from artefacts and background was 99.1%.

3.5.2 Extracted pixel distribution approaches

This section focuses on publications that aimed to develop probabilistic inference over an prior extracted distribution using some type of stochastic-based methodologies. In the literature, both unsupervised (such as $k-$means) and supervised (such as naive Bayes classifier) methods were used.

Govindan *et al.* provided an example of unsupervised learning-based segmentation on Ziehl-Neelsen stained light microscopy images which utilised $k-$means clustering in conjunction with decorrelation stretching (of color bands in an RGB image) to identify areas of interest [91]. Dilating and eroding morphological operators were required to close broken edges in the final segmented images. FDs, eccentricity, and compactness were the feature types utilised to extract contour information. Finally, the candidate contours were classified using an SVM model. Alternatively, Ayas and Ekinci deployed a Random Forest (RF) approach, which is a supervised learning method, to classify each pixel as a possible bacilli area based on local colour distributions [11]. Their method labelled each pixel as either a prospective *Mtb* bacilli pixel or not. Then, each pixel group was rotated, scaled, and centred inside a bounding box before being classified using the RF learning algorithm trained on an image set with manually labelled *Mtb*-containing patches.

Sadaphal *et al.* employed Bayesian segmentation based on a *priori* knowledge of bacterial colour [192]. After the application of morphological operations, a set of shape criteria including axis length, eccentricity and area evaluated whether predicted objects of interest belonged to the bacteria class, were probable bacteria, or were not bacteria. A similar method was described by Khutlang *et al.* in which brightfield microscopy images were partitioned into background and bacterial by two naive Bayes pixel classifiers [115]. Extraction of geometrically transformation-invariant features and optimization by feature subset selection and Fisher transformation were performed on the resulting binary images. The authors compared the outcomes of two object classifiers, NNs and SVMs. Accuracy, sensitivity, and specificity were all reported to be more than 95%. The same group published a second work with a similar two-step approach, but this time segmentation was accomplished using a mixture of Gaussian classifiers [116]. This method worked better for both segmentation and classification. In their second work, overall sensitivity increased by over 2%, while both accuracy and specificity were reduced by more than 4%.

Within this realm, Chang *et al.* used a white top-hat transform and template matching with a Gaussian kernel to binarize images into black background and white regions of interest [34]. Diluting and eroding morphological operators were utilised to close fractured contours. Binarized images were then used for feature extraction using Hu's moments, geometric and photometric features, and histogram of gradients (HOG). Finally, these features were used to classify whether each candidate contour belonged to the bacterial class using SVM. Diaz-Huerta *et al.* also used a Bayesian classifier based on a Gaussian mixture model [60]. Their work consisted only of segmentation, correctly distinguishing bacteria from the background.

A final paper, by Ayma *et al.,* included adaptive signal processing approaches, namely the least mean squares and reduced rank with eigendecomposition algorithms, both of which contain learning parameters for optimization during training [12]. Similar to Diaz-Huerta *et al.* work, this work focussed on segmentation only. Although the authors reported competitive results, a total of 650 images were captured, but only 80 were utilised owing to noise, focus, and stain difficulties. One can argue that the reasons for this dataset reduction are the same ones that inspire automated TB-AI microscopy to help overcome some difficulties in manually examining sputum smears. Section 1.1.1 explains that manual reading of microscopy slides is prone to subjectivity of evaluation due to variability of staining and image generation methods. If only high-quality, noise-free images are used for TB-AI assessment the ability of automation to overcome this issue cannot be assessed and selection bias may occur, limiting comparison with 'real world' studies based on more representative image datasets [73, 74, 114, 153, 241].

3.6 Research utilising DL

This section will present publications with a primary focus on DL methods, including the use of convolutional neural network (CNN) or deep convolutional neural network (DCNN).

CNNs represent a subclass of neural network architectures designed specifically for processing grid-structured data, serving both as feature extractors and classifiers, although these roles are not mutually exclusive. To analyze how a local image region responds to abrupt changes in pixel values, individual kernels or receptive units carry out sequential convolutions across the image. Generally, DL approaches TB diagnosis and *Mtb* detection in two ways. In the first approach an end-to-end customised DCNN architecture is designed for each task. In the second, two-step, approach a separate image processing technique is used for low level feature extraction, e.g. image binarization, contour extraction of objects, or noise removal, and these features are then fed to another CNN technique, either custom or generic. These approaches are not mutually exclusive, and some methods incorporate a combination of both. Table 3.7 summaries result from common evaluation metrics used by all DL papers included in this section.

Table 3.7: Performance metrics used to evaluate DL methods in TB-AI microscopy

Paper	Accuracy	Sensitivity/Recall	Specificity
Lopez *et al.* [134]	N/A	N/A	N/A
Kant & Srivastava [114]	99.80%	83.78%	N/A
Panicker *et al.* [174]	N/A	97.13%	N/A
Samuel & Kanna [63]	95.05%	N/A	N/A
Xiong *et al.* [254]	N/A	97.94%	83.65%
Ahmed *et al.* [4]	96.07%	N/A	N/A
Hu *et al.* [103]	98.40%	98.00%	98.4%
El-Melegy *et al.* [71]	N/A	98.4%	N/A
Mithra & Emmanuel [148]	97.55%	97.86%	98.23%
Serao *et al.* [199]	99.67%	99.98%	99.34%
Swetha *et al.* [263]	N/A	94.7%	99.00%

3.6.1 Custom-made CNN architectures

Like traditional ML approaches, most existing DL work utilize as input pre-cropped FOVs, or tiny patches, from microscopy slides which contain *Mtb* bacteria (or not). For example, Lopez *et al.* proposed a technique for automated classification of brightfield smear microscopy patches employing RGB, R-G, and greyscale patch versions as inputs to a CNN [134]. Xiong *et al.* also described a method using a CNN which was pre-trained on CIFAR-10 (an open-access dataset of 60000 of 32 ×32 pixel images, `https://www.cs.toronto.edu/~kriz/cifar.html`) in order that it could then assess pre-cropped patches from TB microscopy slides of the same size [254]. To improve results, bootstrap training was implemented. Although the authors did not include any further architectural information, the results were promising, with 97.94% sensitivity and 83.65% specificity.

Another method that used manually cropped positive (bacteria containing) and negative patches (void of bacteria) was described by Serrao *et al.* [199]. Each patch binarisation involved the segmentation of background from foreground (*Mtb* containing) regions. Groups of 100 patches were then combined into a 400 × 400 pixel mosaic images. 5000 of these mosaic images were inputs for three CNNs. Results from a range of performance metrics were reported, namely sensitivity of 99.98% and specificity of 99.34% for CNN-1.

All the methods listed so far this section so far have a key restriction in that they were not fully automated, as they focussed on manually cropped patches. Kant & Srivastava also used a patchwise classifier to categorise whether particular patches from TB microscopy images included bacteria or not [114]. However, instead of manually cropping, they used an autofocus method to construct 20 × 20 pixel patches from the whole slide. The CNN used to classify these patches was composed of five convolutional layers and no linear layers. Efforts to automate cropping of regions of interest (either FOVs or smaller patches) from whole TB microscopy slides will be discussed further in Section 3.8.

3.6.2 Gradient-based approaches

The literature includes works that employ a multi-stage approach, initially utilizing gradient-based CV algorithms for image segmentation. This preliminary step offers advantages such as effective noise reduction and computational efficiency, for subsequent processing through CNN. Panicker *et al.,* for instance, utilised the fast nonlocal means method to denoise images from a publicly available FOV dataset, followed by Otsu's threshold to binarize the images into background and foreground [174]. The authors then fed these images into a CNN with five layers and one linear layer for pixel classification. Although their methodology surpassed similar earlier efforts, it was incapable of classifying bacteria with unusual *Mtb* shapes, i.e. anything other than elongated rods. In the work of Mithra *el al.* , the channel area thresholding channel area thresholding (CAT) technique was proposed for bacterial image segmentation of FOV images from the publicly available ZNSM–iDB dataset [148]. Intensity-based local bacilli characteristics were derived utilising a location-oriented histogram and a speeded up robust feature (SURF) algorithm extraction. Deep belief neural networks were used to classify the bacilli items following segmentation stage. In a similar paper, Swetha *et al.* pre-processed brightfield sputum images by noise reduction and intensity modulation prior to a segmentation method which used CAT in addition to extracted features such as HOG and SURF [221]. Classification was performed using a CNN classifier, which classified the input image as mild, moderate, or severe depending on the number of pixels classified belonging to the bacteria class. Although the authors reported sensitivity of 94.7% and specificity of 99.0% for their method, they provided no more information on the architecture of the employed model.

Samuel and Kanna described a method which attempted to automate analysis of whole microscopy slides, without manual cropping [63]. A motorised microscopy stage pre-selected FOVs from the slide. The FOVs were then used to train a DCNN (customised InceptionV3 [222] model) with transfer learning to derive bacteria inference feature maps. Finally feature maps were employed as training data for an SVM to determine whether FOVs from the ZNSM–iDB dataset included *Mtb* cells. Hu *et al.* , published a method providing a classification approach for complete slides [103]. Considering that high resolution digital images of microscopy slides are often several GB in size, the authors developed a dataset creation technique based on non-overlapping subgraph partition. Pre-trained models namely ResNet [98], InceptionV3 [222], and DenseNet [104] with transfer learning were employed to evaluate the performance of their approach. InceptionV3 fared the best, with an error rate of less than 5% when reading a slide for diagnosis. However, when more than one bacillus was present, the subgraph partitioning method sometimes resulted in incorrect bacterial counts.

3.6.3 Employing existing models for Mtb bacteria feature extraction

Methods also leverage pre-trained CNN as foundational elements for feature extraction, either with or without transfer learning. This technique may enhance the discriminative power of traditional ML classifiers, such as SVM, by utilizing the set of feature representations learned by these deep networks. For instance, Ahmed *et al.* presented a method in which they categorised numerous bacteria associated with a variety of diseases [4]. They used InceptionV3 with transfer learning and discarded all fully connected layers, thereby functioning as a feature encoder. Later, the collected features from InceptionV3 were flattened and fed into an SVM classifier. El-Melgey *et al.* presented a work using Faster Region-based convolutional neural networks (Faster-RCNN) to swiftly localise bacteria using ground truth bounding boxes [71]. However, due to the high likelihood of false positives, the authors introduced a second step to determine whether the projected bounded boxes actually belong to the bacterium class. The comparative evaluations were limited to using bounding boxes that could fit only a single bacterium, constraining the analysis to isolated bacteria detection rather than clusters. This represents a limitation of their method, given the biological tendency of *Mtb* cells to clump together.

3.7 Research on Mtb bacteria load quantification

For *Mtb* quantification, some authors manually segmented microscopy images and counted the bacteria present. For example, Sotaquirá *et al.* converted sputum smear images into YCbCr and Lab colour spaces which they subsequently evaluated for their relative difference [212]. Quantification of bacteria was done by computation of the mean size of bacilli, accounting for image resolution and the pixel count of the segmented image. Aside from qualitative and visual results, the authors provided no evaluation metrics. Moreover, the heuristic information at the core of their method was dependent on image dimensions and, as explained in Section 3.3.1, it cannot be assumed that all datasets include images with the same dimensions. Finally, a major limitation of this work was that manual enumeration of bacteria undermined the objective of automating the process.

Nayak *et al.* proposed a technique that employed colour segmentation and colour space transformation [153]. They described a five-step process: i) colour-based segmentation, ii) connected component labelling, iii) size thresholding on the resulting contours, iv) proximity grouping, and v) size constraints. The contours produced by this process were used determine how many bacteria were present. In line with the focus on utilizing pre-trained models for feature extraction, Soans *et al.* proposed a method that segments images and manually counts detected objects of interest using the HSI color model [209]. Given the now segmented image, a knowledge-database was constructed and passed to a decision tree classifier to determine which HSI (Hue, Saturation, Intensity) component values corresponded to the bacterium class. Lastly, proximity groupings and size constraints were used to eliminate false negative objects. Similarly, by thresholding hue range, a hue colour component-based approach was utilised to segment bacilli, and morphological characterisation was employed to determine whether or not bacilli were valid [176]. Through thresholding the area, perimeter, and contour characterizations, other artefacts were eliminated. Using area, perimeter, and shape characteristics, clumps of bacilli were detected. Manual counting occurred followed segmentation of bacilli and bacilli clusters.

Mirtha *et al.* proposed a quantification method comprised of three steps: i) segmentation, ii) feature extraction, and iii) classification [147]. The input sputum smear microscopic image was first subjected to a colour space transformation, followed by thresholding to generate the segmented image. Length, density, area, and histogram characteristics were collected for fuzzy and hyco-entropy-based decision tree (FHDT) based classification, which classified contours as low-bacilli, non-bacilli, and overlapping bacilli. An entropy function, in this case, hyco-entropy-based decision tree (HEDT) of was created for optimum feature selection as input to a decision tree model. The HEDT algorithm's key contribution lies in its ability to simultaneously manage both continuous and discrete variables during the decision tree construction process. Conventional decision trees operate based on information gain, a method well suited for categorical variables but can be unreliable for continuous ones due to their nature of creating discrete partitions of the feature space [196]. The HEDT approach addresses this limitation by incorporating entropy-based techniques specifically designed for handling continuous variables. In addition, a fuzzy classifier is used for classification analysis in order to determine the number of overlapping bacilli. Perhaps the most relevant contribution of this study is that it was the first to automate bacilli counting, compared to previous research that accomplished this step manually.

In one of the most recent publications on the subject, Vente *et al.* proposed a complicated approach for the localisation of bacteria, utilising edge detection, Fourier analysis, and morphological operators, and then calculating the bacterial count in areas of interest using simple regression [241]. The authors reported a 6.5% error in the test set.

My own work on automated bacterial counting will be described in Chapter 5.

3.8 Research on automated FOV acquisition

A recurrent theme in this review is that most 'automated' approaches to TB microscopy require a computer-based model system to be fed with images that are already cropped, either as FOVs or much smaller patches. As a substantial time-consuming component of image analysis is finding FOVs of interest, TB-AI will be most useful if it is able to automate the entire process of microscopy. The few authors to date who have attempted this have deployed one of three strategies: 1) using a microscope with an auto-focus function to select FOVs of interest (via a mathematical algorithm based on colour or image intensity) then applying ML or DL tools to the chosen FOVs, 2) developing an end-to-end model in which an entire slide image is fed into a model which constructs and analyses patches of interest using DL, or 3) simply partitioning a full slide image into smaller sections which are read one-by-one by DL methods, i.e. a filtering stage whereby non-salient FOVs are excluded.

As described in Section 3.5.1, Forero *et al.* used the first strategy in one of the earliest attempts to completely automate analysis of Auramine O stained fluorescence microscopy slides in TB diagnosis [75]. Autofocus of the microscope was accomplished by a two-pass algorithm that determined whether a specific area was void of bacterial content before bringing the image into focus [141]. The initial run of the algorithm analysed slides at three $z-axis$ points to assess if there was sufficient variance to signal the presence of salient content in the field. As *Mtb* bacteria occupy extremely small areas of any image, i.e. most is taken up by background, the experiments by these authors demonstrated that a narrow scanning window (256 × 256 pixels) is required for accurate FOV localization. Four auto-focusing methods were assessed and two (wave and auto-correlation based approaches) produced promising results.

A more recent attempt to make use of autofocus functions is that of Zhai *et al.* [270]. Their work used light rather than fluorescent microscopy images and employed a row-wise scanning strategy followed by three different autofocus measurements: the sum of gray-level differences, the Laplacian, and the Tenengrad function with the Sobel operator. The Sobel operator performed best but the reported accuracy of their overall method was much lower than that of methods which worked on manually cropped FOVs [174, 192, 242]. Kant and Srivastava's used the second strategy to processes entire slides in a bottom-up manner, by aggregating information extracted from small patches [114]. They used a five-layer patch-wise classifier to load each tile from a microscopic slide and a 20 × 20 pixel window which moved through tiles of the slide to assess the presence of bacteria. Although they reported 99.8% accuracy, this is misleading for reasons described in Section 3.4.1. The majority of the area on a microscopy slide is occupied by background, resulting in high accuracy owing to accurate background classification, i.e. in effect, FN errors are deprioritized). When sensitivity and specificity were considered, rates of 83.8% and 67.6% respectively were comparable to or worse than other classification methods discussed in this chapter. Slightly disappointing performance metrics from Forero *et al.* [75], Zhai *et al.* [270] and Kant [114] illustrate an important point. Whilst it is desirable to develop TB-AI microscopy methods which perform automatic selection and detailed analysis of FOVs or patches from whole microscope slides this is an inherently more difficult task than developing models which automate FOV analysis alone.

The third strategy, of simply partitioning large slide images into small sections, which are read one-by-one was used by Hu *et al.* [103] and is also the basis of my own work on whole slide analysis to be discussed in Chapter 6.

3.9 Research on phenotypic characterisation of bacteria

My literature review did not identify any existing studies seeking to automate measurement of phenotypic characteristics of *Mtb* cells. From broader reading on related work, several papers do describe AI-based morphological phenotyping in the context of other bacterial infections, providing proof-of-principle for this strategy. This work will be discussed in detail in Chapter 7, alongside my own experimental approach.

3.10 Discussion

Efforts to automate the analysis of sputum smear microscopy images have gradually advanced over a period over more than twenty years. This chapter has described progress in TB-AI microscopy using traditional ML methods in Section 3.5, DL methods in Section 3.6, and techniques for bacterial quantification in Section 3.7.

However, several obstacles remain to be addressed including a need to standardise image sets and performance metrics for experimental work. Even meticulous slide preparation for smear microscopy can generate images of variable quality with unpredictable artefact and background staining in the sputum matrix. TB-AI work which selects only high-quality images for analysis may report high performance results which would not be replicated in a ‘real-world’ setting. The decision-making process involved in reading stained sputum smears is inherently challenging. If two experienced microscopists were asked to carefully apply manual labels to *Mtb* cells in a series of smears, there would almost certainly be some differences in their labelling. When the same images are read by a computer-assisted system these challenges will remain, and any method of image analysis will always be limited by the standardised quality of the data input. However, TB-AI analysis should, at least, apply the same uniform approach to the reading of ‘difficult’ slides. Development of clear guidelines for creation of image datasets to be used for TB-AI work would be beneficial. Whilst it may be difficult for laboratories around the world to settle on completely unified approaches using identical equipment, closer agreement on essential characteristics of datasets for AI work would remove some of the current variability. Ideally, open-source, standardised and annotated template datasets could be developed across research centres which would save time and resources when developing new methods. WHO, or other international bodies may help to coordinate this effort. ‘Training’ and ‘test’ combinations of standardised and individually created ‘proprietary’ image collections could also be used to study the robustness of new tools, bearing in mind prior experience that methods do not always translate well between datasets [255].

Establishment of standard performance metrics to evaluate new TB-AI microscopy methods would enable researchers to evaluate the effectiveness of their methods across multiple image sets, reducing the need to adjust model parameters and simplifying method comparison. Consensus agreement between groups active in this field, perhaps supported by WHO guidelines, may also be useful here. In this chapter, I have compiled the predominant evaluation metrics utilised for classification, regression, and segmentation. My work shows clear disparity amongst reported metrics for various approaches to the same problem and this hampers researchers’ ability to ascertain the effectiveness of each.

I have observed that conventional ML techniques exhibit a broad spectrum of sensitivity/recall and specificity scores. The successful integration of ML and heuristic knowledge, specifically incorporation of anticipated cell geometric features into the algorithm, is a contributing factor to some methods which show higher sensitivity. However, this approach also presents a challenge as the same factor that enhances the method's ability to detect *Mtb* bacteria, also increases its susceptibility to FP, adversely impacting specificity [270, 116]. Methods that incorporated a preliminary segmentation stage or a hybrid approach, commonly by leveraging CNNs as feature extraction mechanisms and subsequently feeding these feature maps into another classification/regression algorithm like SVMs, consistently attain superior results [103, 199]. In addition, akin to conventional ML methodologies, DL techniques frequently employ amalgamated shape descriptors. These descriptors are generated either through an auxiliary CNN or through image processing algorithms such as HOG, SURF, or CAT [221, 147]. In general, detection of *Mtb* bacteria relies on appearance and shape regardless of whether the approach employs ML or deep learning techniques.

Although the role of sputum smear microscopy in TB treatment monitoring would be enhanced by TB-AI tools to report changes in bacterial load, current literature on *Mtb* bacterial load quantification on microscopy images is too sparse to draw conclusions on the most appropriate methods for this. A few works have utilised a pipeline approach to completely automate of the quantification process, which may involve a segmentation stage. However, their performance remains insufficiently good for clinical use. Similarly, selection of cropped FOVs for analysis from whole slides predominantly remains a manual process and greater effort to automate this is necessary.

An area which requires attention in TB-AI microscopy is the deficiency in explainable artificial intelligence (XAI) techniques. Despite receiving considerable attention for multiple healthcare applications [44], DL algorithms are only slowly being implemented in clinical practice [56]. This is primarily due to the need for the enhanced transparency and interpretability of ML models, particularly in critical applications such as disease diagnosis and treatment. XAI methods strive to enhance transparency and interpretability in the decision-making mechanisms of AI models, often favouring simpler and more comprehensible representations over intricate ones. XAI models should provide justifications for their decisions by emphasising relevant features or patterns in the input data that influence outcomes. This is crucial for establishing a model's trustworthiness [56]. Within the realm of TB research, XAI techniques could be employed to add understanding of the decision-making mechanisms used by models for *Mtb* cell identification and classification. Integration of microscopy and XAI techniques have been explored in related biomedical applications including detection of leukaemia and babesia [3, 68], creating a precedent for future research in the field of TB-AI to incorporate these advantages as well.

Overall, prior progress on TB-AI microscopy using ML and DL tools provides a useful backdrop and incentive to the experimental work which I will now describe in Chapters 4-7.

Chapter Four

Geometry-based features for Mycobacterium detection

Chapter abstract – A critical aspect of any algorithm created to automate analysis of medical images is accurate object detection: in this case, *Mtb* bacteria on sputum smears. Many *Mtb* cells exhibit a consistent morphology characterised by a straight, thin, and elongated form. However, there are exceptions. Clumping or intersection of multiple bacteria can generate unusual structures, e.g. a cross or a crescent shape, and inconsistent uptake or retention of microscopy dyes can alter cell appearance. In practice, a large number of arbitrary bacterial forms are possible. This chapter employs non-learning CV techniques to acquire information about the structure of *Mtb* cells and detect their presence in FOVs from Auramine O stained slides. A Hessian-based ridge detector with a suppression effect and threshold segmentation is used to detect the thin outlines of bacilli-like objects, then a succession of geometric characteristics are used to classify these as *Mtb* or not. A performance metric comprised of Hu moments was used in addition to standard metrics (Jaccard index and SD coefficient) to evaluate the method on two FOVs test sets. Whilst overall performance of this approach compared unfavourably with published methods for the same task, the proposed method included development of an image enhancement technique which will be used in subsequent thesis chapters.

4.1 Chapter introduction

Medical images were added to AI research for the first time in 1970s, an era when AI was trying to demonstrate its usefulness [225]. Early efforts were founded on rule-based, brute-force search algorithms, often known as expert systems [195]. In the 1980s, ML was introduced for diagnostic imaging [225]. Several ML approaches including SVM, Multi-layer perceptron (MLP), and k-means classifiers were used. However, limitations of hardware at that time, such as computational power and memory, thwarted the potential of these ML approaches to solve complex image analysis problems. This was a rate-limiting issue for genuine advancements in AI to study medical images. In parallel, there were efforts to automate TB microscopy using the microscope itself rather than the images; e.g. autofocusing algorithms to automatically concentrate on regions of interest within FOVs [49].

More recently, progress in ML/DL has yielded renewed interest in TB-AI microscopy [142, 211, 218], but most approaches retain heavy hardware requirements, such as large graphical processing unit (GPU) memory, due to the vast calculations needed to complete the work. Consequently, the viability of this strategy is challenged by those who feel that the cost of the activity is excessive for sustainable implementation [238]. This criticism is particularly pertinent in the context of techniques intended to tackle a disease like TB which is mainly prevalent in LMICs. The increased computational and resource demands of DL means that researchers should not automatically resort to its use even if it typically yields superior results; in some cases less computationally intense CV approaches may be able to address some problems more efficiently [159].

There are differences between the overall underlying principles of CV and DL methods. CV techniques such as the scale invariant feature transform (SIFT) algorithm [136], basic colour thresholding, and pixel counting algorithms are not created for specific tasks or designed using specific datasets. Their structure and implementation remains consistent across various applications, but the resulting outcomes may vary [159]. In contrast, DL models use training data during development. The iterative nature of their learning processes endeavour to capture features that optimally align with characteristics of the training data [159]. The success of a network's output is predicated on it being trained on data which are representative of all the input that it will subsequently receive. Whilst the ability to refine DL model parameters using training data can improve performance, it also creates the risk of model overfitting, a widely recognised and substantial hurdle within the domain of ML [88]. For example, if a model for TB-AI microscopy is too carefully trained on a single image set, this can constrains its capacity for generalisation to new images generated in a different setting under different conditions. In contrast, non-learning CV techniques which do not incorporate a data-driven learning process offer a more consistent approach, yielding predictable outcomes, without the concern of overfitting.

Training datasets for supervised ML and DL models is laborious to assemble; for classification and segmentation approaches to TB-AI microscopy, *Mtb* cells within a dataset of images must be manually annotated by an expert microscopist before use in model training. As large training datasets are often necessary, the process of annotation can be prohibitively time consuming, discouraging use of DL for many clinical applications [247]. Work with non-learning CV algorithms can begin without the need for supervised learning and renders annotations unnecessary at that stage. Of course, evaluating the performance of any method (CV or DL) eventually necessitates a comparison with ground-truth labels, meaning that annotated images are ultimately required. However, the need for ground truth data in both learning and testing phases of supervised ML/DL methods demands that a higher quantity of labels are available, and that they are accessible much earlier in the experimental process.

These observations have relevance to the chronological workplan of this thesis. The early phase of this project coincided directly with the onset of the Covid-19 pandemic, when the clinical supervisor needed for dataset annotation was assigned to emergency hospital-based duties, delaying the availability of a formal image training set for several months. This potential obstacle created a research opportunity to explore CV-based approaches to TB-AI microscopy and evaluation of the resulting methods was completed once annotated ground truth images were available.

This chapter sets out with two main goals:

- To establish whether automated *Mtb* detection can be done using only non-learning CV techniques, without requiring supervised ML or large annotated datasets for model development.

- If the first goal is achieved, to assess whether regions of interest for bacterial detection can be used to estimate the total number of bacteria in a given FOV, providing an introduction to bacterial load estimation. This is considered important, given the lack of data on methods for *Mtb* load quantification revealed in Chapter 3.

Overall, in this chapter, I aim to establish whether non-learning CV methods can reliably detect and count *Mtb* cells without extensive training from ground truth labels, leveraging the inherent interpretability and visualization capabilities of classical image processing.

4.2 Methodology

Here I will provide some background on the principles and mathematical tools underpinning my proposed method for this chapter, before describing the proposed method directly.

4.2.1 Segmentation in digital image processing

As described in Sections 3.5 to 3.7, before classifying or counting bacteria, many of the more effective TB-AI microscopy methods from prior work began with a segmentation stage. Segmentation involves subdividing images into smaller sections which are easier to interpret, because elements of interest are brought to the fore and everything else is excluded as background [18, 201]. This requires examination of differences or similarities between neighbouring pixels in each image according to predefined properties, e.g. colour or intensity. Most segmentation techniques are either boundary based (dependent on finding sudden changes in pixel properties which represent edges or ridges) or region-based (reliant on finding groups of pixels with common properties across part of an image, e.g, using clustering [261] or thresholding).

Mtb bacilli are mostly straight, thin, and elongated in shape so edge-based segmentation to identify the cell perimeter is useful. However, in cases of noisy images, such as those with dye spillage or background artefacts on a microscopy slide, the literature suggests employing region-based segmentation as a more effective approach [259]. The method developed in this chapter will include elements of both approaches. During edge detection, neighbouring pixels with similar predefined properties are linked together to form a contour. The objective of edge detection is to identify significant transitions in image intensity, which often correspond to object boundaries or other important structural elements in the image [201, 136, 31]. A sudden and significant change in the intensity values of neighbouring pixels can indicate the existence of an edge. Variation in intensity between adjacent pixels is not limited to a specific direction and the directional change that exhibits the greatest magnitude is given as a vector, i.e. the image gradient [31].

The definition of an edge in image processing is a localised and significant variation in intensity or colour, such as a traverse of contrasting dark and light values that demarcate a transition between different regions or objects. In contrast, ridges are thinner lines in comparison to edges and have pixel intensities or color values that are either darker or lighter than those of their adjacent pixels. Edges are often spotted using a first order derivative operator, such as the Laplace operator used in the Canny edge detector [31]. Calculations for locating ridges, on the other hand, need second order derivatives, which indicate how much the gradient changes [178, 38]. The difference between an edge and a ridge can be visualised in Figure 4.1 [1]. The gradient is a fundamental concept in the domain of edge and ridge detection, as it quantifies the magnitude and direction of the rate of change of intensity or other image attributes.

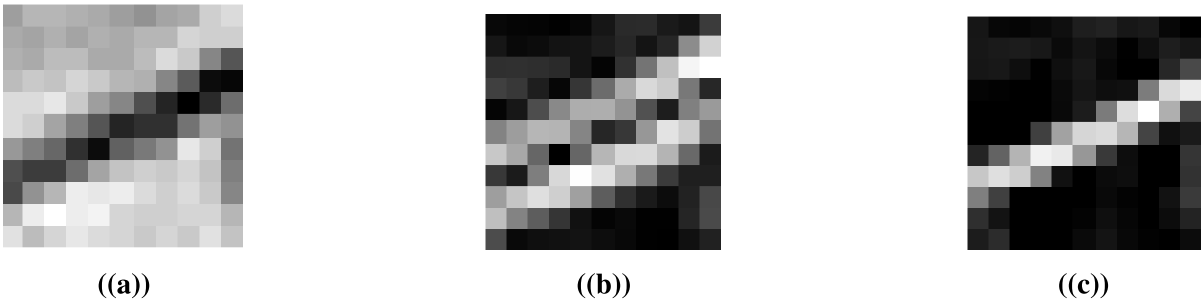


Figure 4.1: Figure (a) shows an arbitrary patch from a greyscale image belonging to a thin line. In this case the background is bright while the lines are darker. Figure (b) shows the same patch with the same line represented as an edge. Lastly, figure (c) has the original line represented as a ridge.

4.2.2 Computing image gradients

A two-dimensional matrix may be used to represent a binary or greyscale image. Each x and y point in the matrix represents the pixel's intensity value. Binary format images have pixels values typically 0 or 1 while greyscale images range from 0 to 255, low to high — dark to bright. As mentioned earlier, in order to differentiate the background from an object within an image using boundary-based segmentation, it is necessary to locate edges or ridges at specific locations. The mathematical functions commonly used for this will now be described.

The derivative of a function measures the function's rate of change in relation to its independent variable. The partial derivative of a multivariable function quantifies the rate at which the function changes with respect to each individual independent variable. Quantifying each partial derivative is done by measuring the extent to which the value of a function changes when each independent variable is altered, while holding all other variables constant. Symbols such as ∂ are commonly employed to represent partial derivatives, which are utilised to calculate of the gradient of a multivariable function. The gradient is a vector composed of the partial derivatives of a function with respect to each independent variable. This gradient pertains to both the direction and magnitude of the steepest ascent or descent of the function. The gradient of an N-variable function at each point is an N-D vector with components obtained by the N-direction derivatives, in this case a 2-variable function, $f(x, y)$ [201, 87], given by:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \tag{4.1}$$

In the physical world, where light intensity exhibits continuous variation throughout a given scene, it is plausible to conceptualise images as continuous functions. However, in digital image processing and representation, it is generally accepted to define images as discrete signals. In order to perform image processing on a computer, it is necessary to sample the image at discrete positions along both the horizontal and vertical axes. The process involves partitioning the continuous image into a matrix of individual pixels. Each pixel serves as a discrete sample of the intensity at a particular location, typically achieved by employing a two-dimensional matrix. Consequently, image derivatives can be approximated by finite differences. Some finite difference techniques are forward, backward, or central and the choice depends on the specific application and the desired accuracy of the derivative approximation. To determine the potential variation in intensity between pixels, consider the image function $f(x, y)$ where x and y denote the spatial coordinates of the pixels under consideration, while i and j are indices which represent discrete positions within the image:

$$\begin{aligned} x\ direction &\Rightarrow f'(x,y) \approx \frac{f(x_i+h,y_j)-f(x_i-h,y_j)}{2h} \\ y\ direction &\Rightarrow f'(x,y) \approx \frac{f(x_i,y_j+k)-f(x_i,y_j+k)}{2k} \end{aligned} \tag{4.2}$$

h and k are the step sizes to define the interval between the points at which the function values are sampled to estimate the derivative; smaller h or k would allow more precise estimation but requires more computational effort while larger h or k would lead to less accurate approximation but less computational effort. Utilizing small convolution filters of size 2×2 or 3×3, such as the Sobel [210], Roberts [228], and Prewitt [178] operators, it is also possible to calculate image derivatives. Using these filters is more precise and can also filter (or smooth) an image and sharpen its edges simultaneously. The output of convolving filters (also known as kernels) over an image is sometimes referred to as a feature map; notably, the core functionality of CNNs involves kernels traversing over an image to produce what are commonly termed as receptive fields. For instance, given an image I and the Sobel kernel K, its first-order derivatives may be calculated as follows:

$$\begin{aligned} x\ direction &\Rightarrow f'(x) \approx I * K_{Sobel} \\ y\ direction &\Rightarrow f'(y) \approx I * K_{Sobel^T} \end{aligned} \tag{4.3}$$

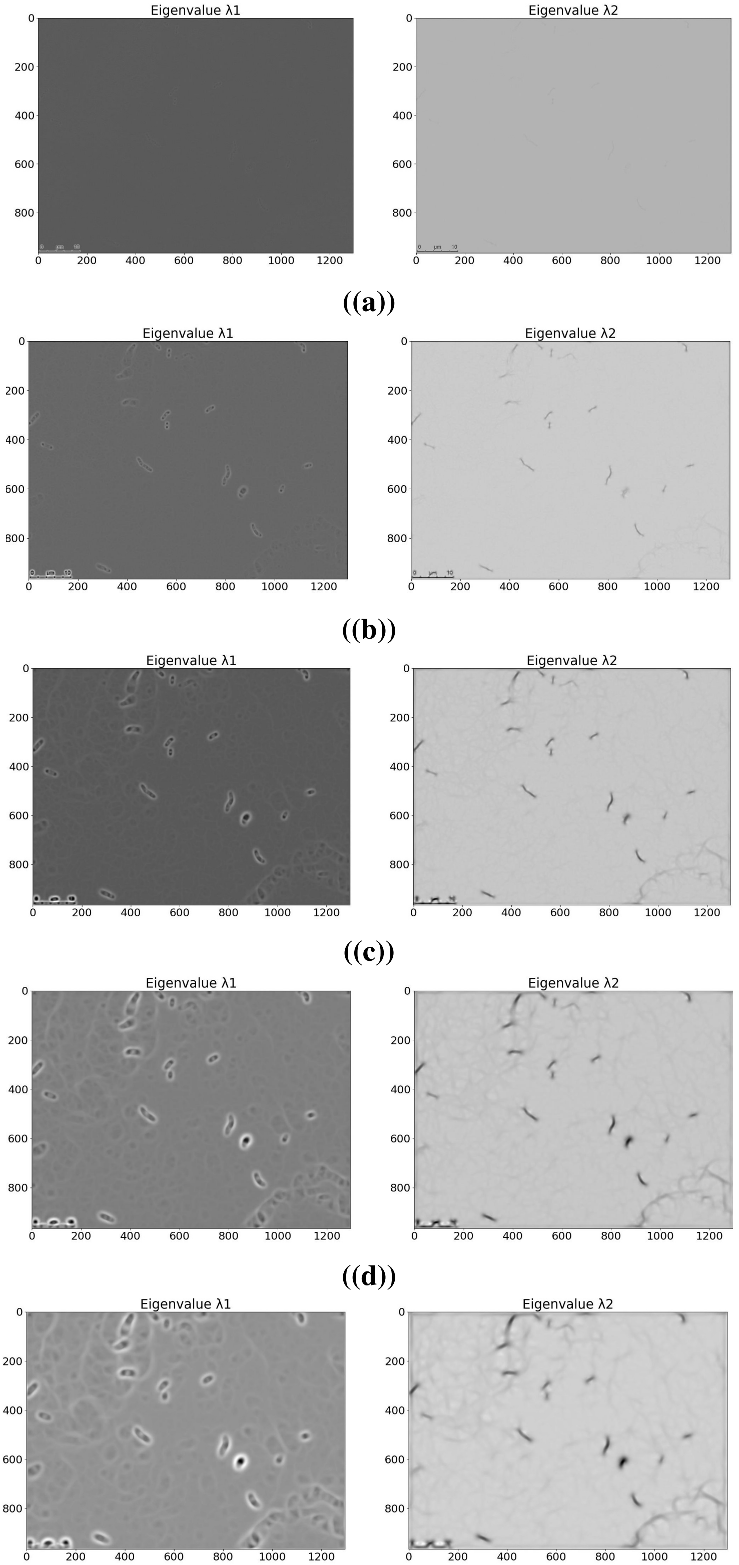
A wider kernel will often provide a better approximation of the derivative, with derivatives of Gaussian functions serving as an example of such filters [29]. For specific applications that require the removal of high-frequency noise, the image undergoes pre-processing via blurring prior to the convolution operation, which in turn emphasizes edges or ridges with strong intensity. For example, within TB microscopy datasets, dye spilling will induce a shift in pixel intensity, from a dark background to a lighter shade. As a low pass filter, the Gaussian filter has the added feature of including this behaviour within the convolution.

4.2.2.1 Hessian matrix, eigendecomposition and principal curvature

The Hessian matrix is a square matrix of second-order partial derivatives and second-order cross partial derivatives of a scalar-valued function. It may be used to characterise the local curvature of a function with several variables, i.e. critical points. Its mathematical properties (second partial derivatives), along with observations of bacillary structure, considerably support the adoption of a Hessian-based ridge detector for segmentation of TB-AI images [55]. Notably, the eigendecomposition of the Hessian matrix allows for a differentiation between different kinds of local image behaviour, leading to a straightforward process of distinguishing between blob-like structures, uniform regions, and elongated structures which may be bacilli. In particular, consider the Hessian matrix with the size of 2×2 pixels, at the scale σ and the image locus P:

$$H = \begin{bmatrix} L_{xx}(P,\ \sigma) & L_{xy}(P,\ \sigma) \\ L_{yx}(P,\ \sigma) & L_{yy}(P,\ \sigma) \end{bmatrix} \tag{4.4}$$

where L_{xx} (P,σ) is the convolution of the second-order derivative of a Gaussian $\frac{\partial^2}{\partial^2 x} g(P,\sigma)$ with the second derivative of an input image at point P, likewise for L_{xy} while L_{yy} are the cross partial (first-order) derivatives [189, 38]. The scale of the Hessian is governed by the value of σ; the smaller its value the finer, (i.e. more local) the scale, and conversely, the greater its value, the coarser (i.e. more global) the scale. The determination of the value of this parameter can be seen as a trade-off emerging from the observation that *Mtb* cells in fluorescence images can be distinguished from the remainder of the image content both by their characteristic shape and apparent brightness. In some instances, the shape is less informative, e.g. clumping or overlap of bacilli may distort or mask their individual shape leaving brightness as the primary cue, whilst in other instances, the bacteria do not exhibit the expected brightness, e.g. if there is poor fluorescence dye uptake or loss of acid-fastness by individual cells, so the shape becomes the most useful cue. Both aspects of appearance need to be considered for maximum robustness, which affects the choice of σ. Figure 4.2 offers a visualisation of the effect applied on the image by σ given a range of values.



((a))

((b))

((c))

((d))

((e))

Figure 4.2: : Figures (a)-(e) show different values for the σ parameter in the Gaussian kernel and their effect on a TB-AI microscopy FOV. The σ values used were odd numbers in the range of 1-9, showing their effect from (a) to (e) respectively. While smaller numbers (1 in (a) and 3 in (b)) did not obtain meaningful geometric information for bacteria, larger numbers (7 in (d) and 9 in (e)) obtained excessive information: because the image's smoothing impact was greater than it should have been, non-bacterial objects, such as artefacts, were emphasised, resulting in an increase in total image noise.

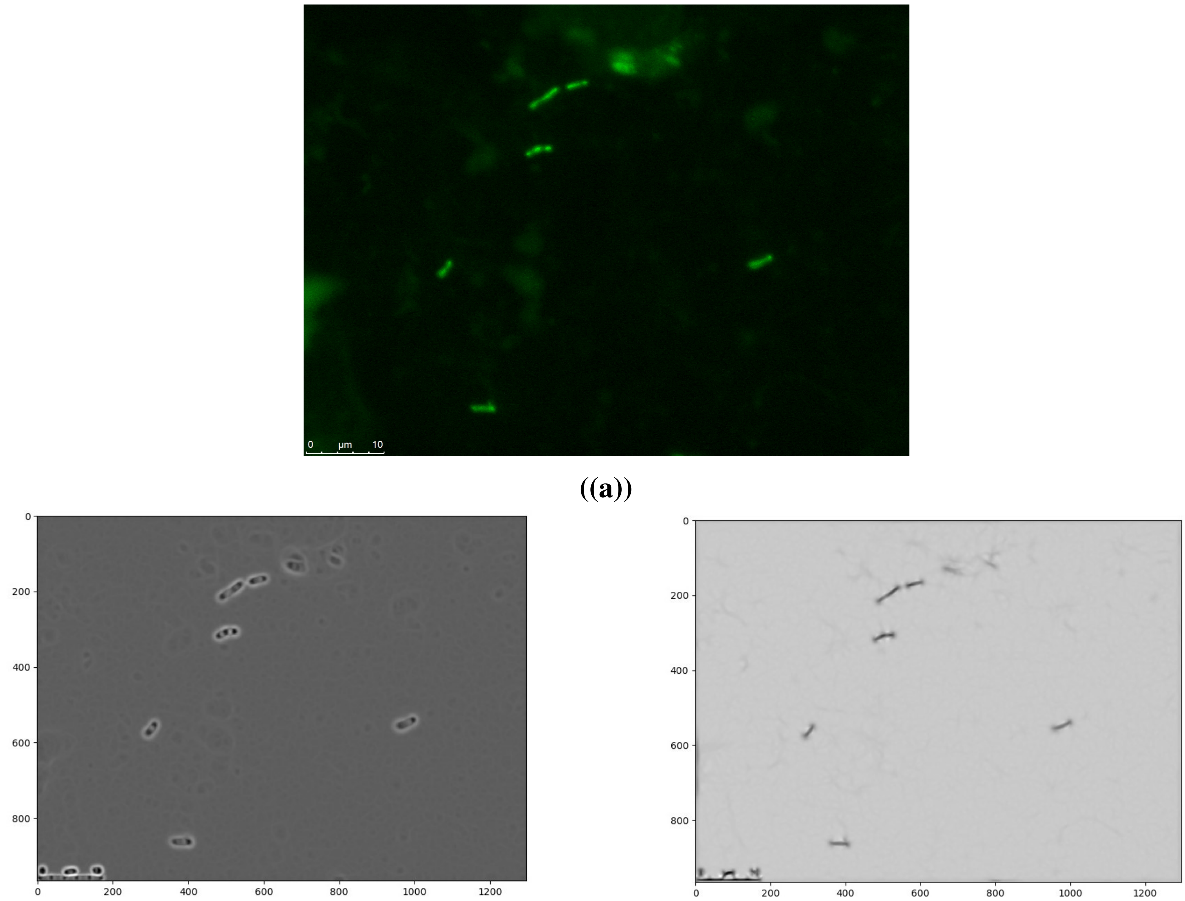
4.2.3 Proposed method

In the proposed method, Hessian matrices are computed for all image loci in the dataset, then used to extract the pseudo-likelihood of each pixel being incident on an *Mtb* bacterium. Recall that the Hessian is informative regarding the nature of local appearance variation in an image [119]. Specifically, since bacilli form elongated structures, loci that demonstrate large change in one principal direction (perpendicular to a bacterium) and minimal change in the other (along a bacterium) may be easily detected using the respective Hessian matrix eigenvalues [82]. To create an enhanced image (in the context of the end goal), each pixel in the original image is replaced with the value of the higher magnitude value and lower magnitude value of the Hessian eigenvalue computed at the locus, referred to as λ_1 and λ_2 respectively; see Figure 4.3. After an iterative experimental process of manual trial and error on a small set of typical images from the Image Test Set 1 (see section for details of image sets 4.3.2) the value of σ was chosen at 5 (this may be compared with the approach to σ choice in related work [38, 70, 189]).

4.2.3.1 Segmenting images based on the eigenvalue ratio

Eigenvalues of the computed Hessian matrix play a crucial role in various image detection and processing applications. They provide information about the curvature of image intensity levels at each pixel and can be used to determine if a pixel belongs to the sought-after object in the image. However, their use is contextual and application dependent. For example, if both eigenvalues of the Hessian matrix at a particular place are strong positive values, this indicates that the surface is concave at that point and suggests the existence of a dark blob [189]. Instead of manually analysing the eigenvalues, and in order to approach the task in an unsupervised manner (i.e. no annotation labels required on the images), the ratio between the two eigenvalues ($\lambda_2 \div \lambda_1$) is calculated. When both eigenvalues at a pixel have similar scalar value, their division yields a low number, signifying that the pixel primarily represents the background. If the result of the ratio between the eigenvalues is very close to 0, it indicates that the pixel in question belongs to the background. Conversely, if the ratio is large, this suggests the presence of a structural element of interest.

The eigenvalue ratio from preliminary sample of images from Image Set 1 has a wide range, from 5.8e−8 to 1.0e+8. The wide range of values in the ratio matrix renders it unsuitable for direct detection tasks; further processing is essential to mitigate the excessive noise commonly induced by division operations. It is unclear at which point the ratio starts to represent detection of useful structures and using a colourmap to create an image from this matrix yields a black image filled with scattered white dots. Therefore, a suppression effect should be applied to the ratio matrix, setting at an appropriate threshold value to remove high but very infrequent ratios which distort



((a))

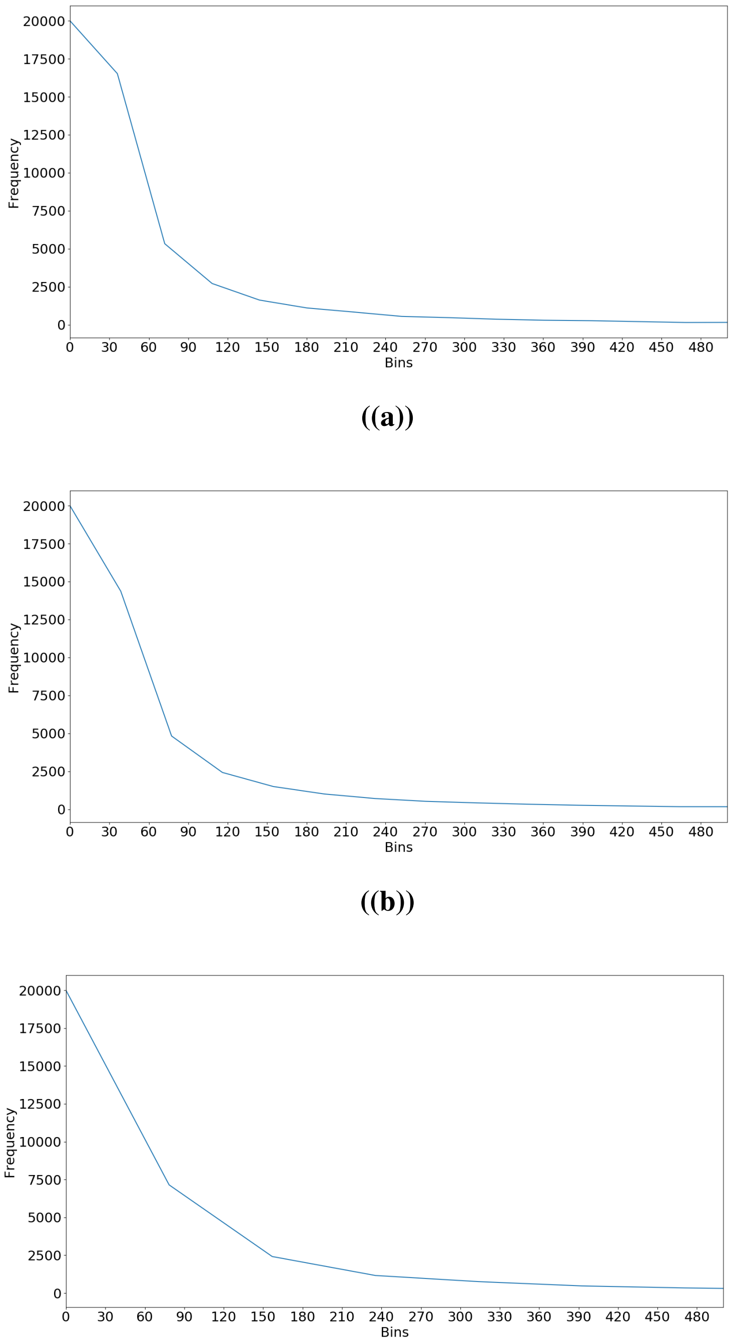
((b))

((c))

Figure 4.3: (a) Example of the typical fluorescence microscopic image used, and (b, c) the corresponding output using the eigenvalues in higher-magnitude (λ_1) and lower-magnitude (λ_2) respectively in conjunction with a greyscale colourmap to display their matrix values as an image.

interpretation of the entire image. If the ratio value is more than the threshold value, then it should be suppressed to the threshold.

To choose an appropriate threshold value for the suppression effect, histograms of 100 images from Image Set 1 are used to determine which values of the ratio matrix have consistently low frequency: consistently high frequencies are expected for background information (or information close to the background), whereas ratios reflecting a structure arise less often, and noise least frequently. These histograms are built using 12000 bins of width 30, which is sufficient to display the frequency shift and has the same range as the ratios. Figure 4.4 shows typical histograms from three images. From this figure, the frequency of observed ratios drops very quickly after the fourth bin (so ratios >480 are not shown). Setting the suppression threshold value within bins 0-30 and 30-60 removes too much information from the image, whilst setting a threshold >90 retains excessive noise. The key dynamic range in all histograms is the third bin, so the threshold is set at 75, the midpoint of that bin.



((a))

((b))

((c))

Figure 4.4: (a)-(c) are three examples of eigenvalue ratio histograms from individual images in the Test 1 dataset. All exhibit an abrupt shifts until the third bin, between the numbers 60 and 90 followed by consistently low frequencies for all input ratios. The suppression threshold is set at 75 (the mid-point of that bin).

Table 4.1: Case Number 1 relates to objects that exhibit clear visibility with distinct pixel intensity and sharpness, consistent with stained *Mtb* cells. Case Number 2 relates to objects with diminished pixel intensity and reduced clarity, but which remain distinguishable from the background as possible *Mtb* cells. Case Number 3 relates case, is for objects with inconsistent pixel intensity in specific areas of their shape (perhaps due to variable fluorescent dye update) but which, overall, have features that are still compatible with *Mtb* cell morphology. Visual examples can be seen in Appendix A.2

	Eigenvalue λ_1		Eigenvalue λ_2	
Case Number	Type	Range	Type	Range
1	High -ve	$(-1.0e-10, -2.1e-10)$	Low -ve	$(-2.5e-10, -5.0e-10)$
2	Low -ve	$(-1.5e-10,; -6.9e-11)$	High -ve	$(-7.0e-10, -1.2e-11)$
3	Low +ve	$(1.0e-10, 8.9e-11)$	Low -ve	$(-2.5e-10, -5.0e-10)$

4.2.3.2 Threshold-based segmentation

The application of this suppression effect results in a binarized image that allows for the detection of *Mtb* cells, however numerous artefacts are also present. It facilitates the creation of a segmented image with potential bacilli and a range of artefacts with varying morphology and dimensions. To minimise these artefacts, threshold segmentation is used, which is perhaps one of the most widely used segmentation techniques [261]. It is a straightforward technique that splits a greyscale image data according on the pixel value of each target [261]. There are two approaches to threshold segmentation: setting a global target for the entire image set, or a local target with numerous threshold values for parts of the data with distinct requirements for each [261]. For this task, a global threshold is employed to satisfy a universal requirement for all images. Similar to the work of Rudzki [189], both eigenvalues are examined and considered to generate three criteria that indicate the presence of possible bacteria. At this stage, the objective is not to definitely detect and classify objects as *Mtb* cells, but rather to focus on specifically identifying and isolating elements whose geometry requires more careful examination, while excluding all other components of the images. For this step, any point that does not meet the requirements indicated in Table 4.1 is assigned a value of 0 in the ratio matrix, i.e. it becomes background and is ignored.

In some instances, the region inside an Auramine O stained *Mtb* cell is not of uniform brightness. Therefore, when the above approaches to image processing are employed, the apparent shape of certain bacteria is altered. Hysteresis is used to enhance individual pixels in order to build the complete shape of the bacterium. Hysteresis is a technique utilized in procedures like the Canny edge detector, in which two thresholds, one low and one high in value, are employed for the purpose of distinguishing between weak and strong edges in an image [31]. Each pixel whose intensity value exceeds the low threshold but does not reach the high threshold is utilized to verify whether a neighbouring pixel satisfies the high criterion within an 8-connectivity neighbourhood [31]. If the condition is met, the pixel with low intensity is enhanced; if the condition is not met, it is considered to be unrelated to an edge or ridge, and therefore excluded. In this work the low value is set to 0.1 and the high value to 2.

The hysteresis mechanism does not guarantee restoration of the contours to their exact original form, i.e., the shape of a bacterium, and relying solely on hysteresis often proves inadequate to connect all the contour pixels. Similar threshold segmentation approaches employed for TB bacteria have encountered similar obstacles in previous methods [76]. These challenges stem from the operator proficiency and unpredictable dye uptake during the preparation process of microscope slides. To address this limitation, the application of dilation and erosion techniques was deemed essential. These fundamental morphological operations systematically enlarge the boundaries of foreground pixel regions (objects), thereby mitigating the identified constraint. Both image dilation and erosion are applied using a square kernel (20×20 in this case). Again, the kernel's size is dependent on the application. However, smaller kernels may compromise the intended effect, as they are less effective in reconstructing contour boundaries.

4.2.3.3 Geometry based reduction of false positive detections

Whilst reinforcing the ridges and repairing the damaged links within likely bacterial objects is useful, this process has the same effect on the remaining artefacts with an unwanted consequence of increasing the possibility that they will be mistaken for bacteria.

To characterise the appearance of bacteria more precisely, a set of geometric features are compiled. As bacteria commonly exhibit an elongated morphology without complicated additional structures, one approach would be to determine whether a particular object had a rod-like form. However, this criterion is often insufficient in cases where bacteria are clumped or overlap one another. Consequently, attempting to determine whether contours form a "rod" shape by calculating the slope of the line connecting two points using the arctan function (inverse tangent) is ineffective. Another approach, encompasses a array of shapes with a broader spectrum of geometric criteria, including perimeter, area, and form related to possible bacillary shapes. For every contour detected within an image, critical points that outline the boundary of the object are recorded. Subsequently, the arc length and area of each distinct object are computed using the contour outlines. This computation involves evaluating the distances between successive contour points, and follows a process akin to the shoelace formula [30] to accurately determine the area.

The Douglas-Peucker [67] algorithm is utilized to determine the approximate form of the contour. This algorithm simplifies the contour by iteratively identifying points that significantly deviate from a straight-line approximation. The degree of deviation is gauged by the epsilon (ϵ) parameter, which is typically set to a fraction of the arc length (e.g. × 0.1). The choice of ϵ determines the level of simplification. A smaller ϵ will result in a more accurate approximation but may retain more points, while a larger ϵ will yield a more aggressive simplification with fewer points. Points exceeding this threshold are considered substantial deviations and are preserved in the simplified representation, while those falling within the threshold are gradually eliminated to achieve a reduced yet visually accurate approximation. This strategy allows the contour to be efficiently represented with fewer points, proving beneficial for scenarios where data size and computational complexity are critical factors. Subsequently, four distinct criteria for bacteria classification are compiled as shown in Table 4.2, in order to eliminate artefacts from the final segmented binary image.

Table 4.2: Geometric characteristics of objects classified as Mtb cells in final binarised image

Case number	Description
1	Area between 100 and 2050 (in pixels)
2	Perimeter between 70 and 800 (in pixels)
3	Length of approximate shape between 9 and 20
4	Number of contour points above 55

4.3 Results

This section describes an empirical evaluation of the proposed method using real-world TB fluorescence microscopy data. The datasets used in this chapter are also used for experimental work in Chapters 5– 7. Their acquisition and structure are explained first, followed by a series of statistical analyses and comparisons between expected and predicted segmented images.

4.3.1 Environmental setup and quality control

The implementation of the proposed method was performed using Pytorch, as were all experiments presented in this thesis that require heavy use of the GPU component. The GPU used for all DL experimental setups was Nvidia GeForce GTX TITAN X. Significant libraries and frameworks employed include skimage, Pytorch, sklearn, and scipy. The experiments in subsequent chapters predominantly feature CNN developed from the ground up utilizing Pytorch's comprehensive features. Furthermore, ML techniques were executed using the sklearn library.

For quality assurance during experiment execution, each segment of the experiment was maintained in separate files for enhanced modularity and fault tolerance, facilitating easier bug detection. Furthermore, each input image and its corresponding ground truth were assigned matching index numbers to ensure accurate pairings, thereby bolstering validation efforts.

4.3.2 Dataset acquisition

Two datasets of images were used.

Image Set 1 comprised microscopy images obtained from a clinical cohort study in Tanzania which was completed between February 2017 and March 2018. Details of clinical and microbiological aspects of that research have been published separately [149]. In brief, 46 adults (40 new and 6 previously treated), aged $\geq$ 18 year with sputum smear-positive pulmonary TB were recruited at clinical facilities affiliated to NIMR-Mbeya Medical Research Centre (NIMR-MMRC) and followed up until the end of a 6-month course of standard TB treatment, between February 2017 and March 2018. Smears on microscopy slides were prepared from sputum samples collected pre-treatment and after 2 weeks, 2 months and 5-6 months of therapy. These were stained according to standard Auramine O LipidTox Red (LTR) protocols and viewed at ×1000 using an oil immersion lens of a Leica DM5500 microscope with a DFC 300G camera attachment. Paired FOVs containing *Mtb* were photographed at manual microscopy, using an N3 filter cube (excitation and emission spectra of 546/12 and 600/40nm) to assess Auramine O staining and a TX2 filter cube to assess LTR staining (excitation and emission spectra of 560/40 and 645/75nm). The full microbiology Standard Operating Procedure for microscopy is included as Appendix A.1. A total of 230 slides were examined, and an average of 30 pairs of Auramine O and LTR FOV images were generated for each AFB-positive slide. As the work of this chapter is focussed on bacterial detection than lipid-based cell phenotyping, only Auramine O stained FOVs were used at this point.

500 Auramine O stained FOVs were selected at random across all timepoints of sample collection. In order to allow create a ground truth for evaluation of the segmentation method developed in Section 4.2.3, these were annotated by a microscopist who carefully drew around the perimeter of each *Mtb* cell. 300 of the annotated FOVs were used for performance evalulation. Although annotated images were not required (or available) during the period of method development for this chapter, a sub-set of images was informally used for some steps in iterative method optimisation (e.g, setting the σ value and the suppression effect for eigenvalue ratios based on individual image histograms).

Image Set 2 was derived from a separate sputum sample set [2]. Two positive sputum smears from a previous clinical trial of TB therapy were stained using the same Auramine O LTR protocol as shown in Appendix A.1, but the work was done in a different laboratory by a different microbiologist [2]. Instead of field-by-field manual microscopy to photograph relevant FOVs, the whole slides were imaged using a Zeiss Axioscan Z1 scanner with Zeiss Zen 3.1 software. Optical magnification of 40 was used, supplemented by digital zoom on the scanner. Whilst the main deployment of these images was for work to be described in Chapter 6, 150 FOVs containing AFBs were annotated and used as a second evaluation dataset in this chapter.

4.3.3 Evaluation of image similarity and shape characterisation

As the majority of a FOV or a slide is background, pixel-wise classification accuracy may lead to misleading performance evaluation and is inappropriate for the reasons discussed in Section 3.4.3. Recall, high pixel-wise accuracy can be attained by correctly categorizing the abundant background regions, even if the model falls short in accurately detecting the more significant foreground bacteria. In the context of *Mtb* detection, accurate classification of background pixels is less relevant than the precise localization of bacteria. Therefore, it is preferable to quantify the similarity in form between the contours of pairs of binary images.

Within Image Sets 1 and 2, annotated ground-truth binarised FOVs were paired with their corresponding model-predicted binarised images. As described in Section 4.1, CV-based techniques for image segmentation do not require a training dataset for the model learning phase, so annotated ground truth images were only available for performance evaluation.

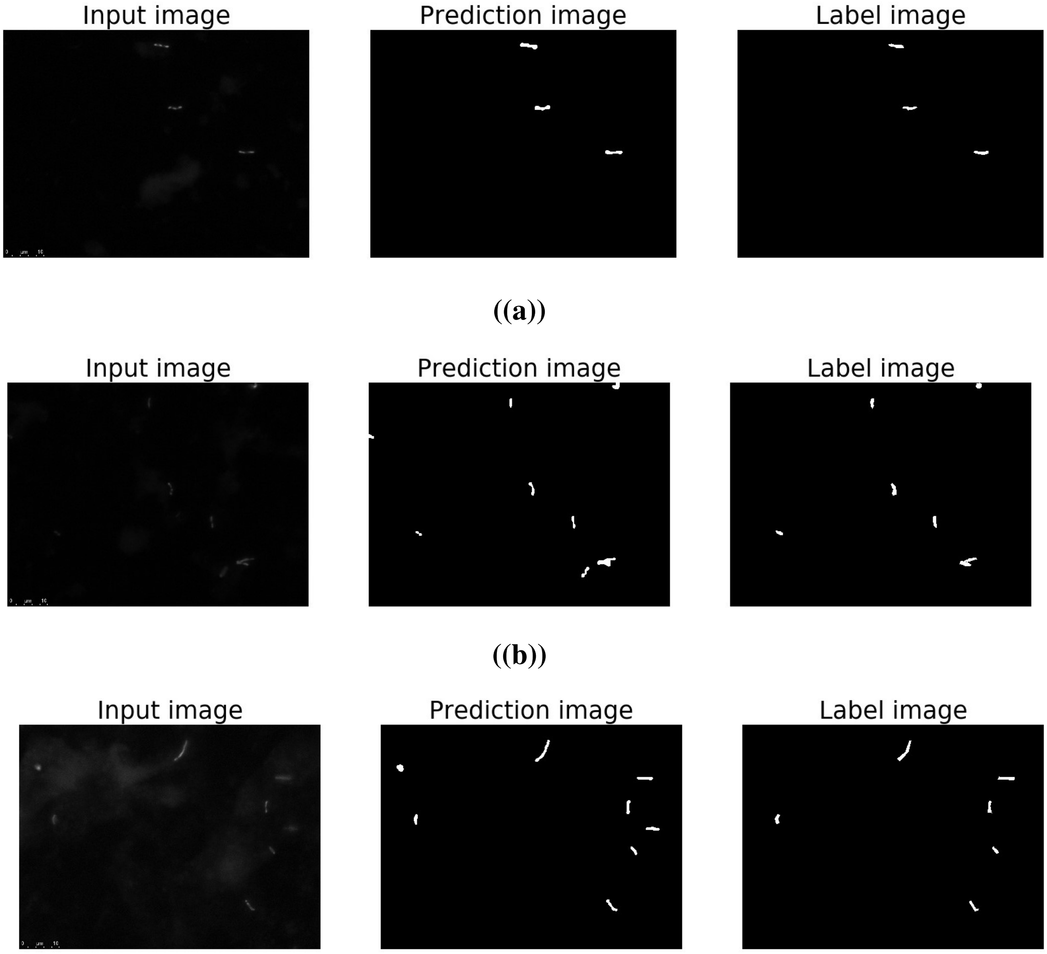
Two approaches were taken for performance evaluation. Firstly, Hu moments (or Hu moment invariants) were used. These are a collection of seven values derived from image transformation invariant central moments [105]. The first six moments are invariant under translation, scaling, rotation, and reflection [105]. At the seventh order, the sign changes for image reflection [105]. In essence, Hu moments extract higher-level features from the image, enabling robust and distinctive characterization of shapes and pattern based on properties such as area, centroid, and spatial distribution of intensity. If a given pair of contours have the same form in both ground truth and predicted images, they should exhibit a low distance value between their Hu moments, regardless of the image size and magnification levels. Binary images, based on the contours of bacteria delineated within each ground truth and predicted FOV were used to compute raw image moments before calculating the seven dimensional vector of the Hu moments. In computing the distance (Δ) between feature vectors, the L_1 norm or absolute error is used as the distance function. The overall procedure is described as:

$$\Delta(P,G) = \sum_{i=0}^{6} \left| H_i^P - H_i^G \right| \tag{4.5}$$

Table 4.3: Performance evaluation metrics for CV algorithm across two image sets

Metric	Image Set 1	Image Set 2
Mean	22.04	20.09
Standard deviation	0.81	2.85
25th percentile	21.60	18.00
75th percentile	22.90	22.81
L_1 norm	6609.15	3014.25
L_∞ norm	23.82	24.94
Jaccard index	73%	69%
SD	81%	78%

where H_i^P is the Hu moment vector from the predicted contour and likewise H_i^G from the ground truth contour. Note that the last value of the vector is ignored as it is simply the sign change for image reflection and therefore not applicable in this case. Table 4.3 displays the aggregate data findings for the two test sets and Figure 4.5 summarizes results. Overall, for bacillary shapes which are generally not very complicated the L_1 and L_∞ measurements do not suggest satisfactory performance of the method in detecting *Mtb* cells.



((a))

((b))

((c))

Figure 4.5: Examples (a-c) illustrating the performance of the suggested method on a succession of original — predicted — ground truth. Note that the predicted image is the end result after the original image has passed through all of the proposed method's components.

The second approach to performance evaluation utilised the Jaccard Index and SD. These metrics served a dual purpose in this context. Firstly, they addressed the need for standardized evaluation metrics, as discussed in Chapter 3. Secondly, whilst the inclusion of Hu moments sought to gauge the algorithm's ability to accurately describe the shape of a bacilli, the Jaccard Index and SD fulfilled the role of assessing how effectively the algorithm localizes bacilli at their precise locations. Results are also shown in Table 4.3. For both image sets, results were worse than prior methods reporting the same standard segmentation metrics, as described in Chapter 3.

The overall results offer further interesting insights. Specifically, the Jaccard Index and SD exhibit higher values for Image Test 1, indicating that the FOVs in this set are comparatively easier to interpret and to locate bacteria within. Conversely, the metrics related to shape characterisation

(Hu moments) perform better in Image Set 2, suggesting that while it may be more challenging to detect bacteria in this set, the bacilli tend to exhibit simpler, less clumped and easier to outline shapes.

4.3.4 Estimation of bacterial number per FOV

Although the second aim of this chapter was to estimate the number of cells in each FOV, performance of the method developed so far was deemed unsatisfactory to proceed with this step. Instead, different approaches to bacterial detection were investigated, using DL, and these will be discussed in Chapter 5.

4.4 Discussion

This chapter demonstrated the feasibility of geometry-based Mtb detection through the exclusive use of CV methods. Initial assessment of the results suggests unsatisfactory performance in bacterial detection in comparison to other methods from the literature. However, similar approaches (i.e. where the presence of a classifier or learning from training data during model development is absent) have sometimes reported better results using datasets comprised of selected image patches containing individual bacteria [76, 137]. As mentioned in Chapter 3, this practice can introduce a selection bias to the research, so that reported results might not accurately reflect (and may favourably over-claim) how a given technique would perform in real-world assessment of unselected TB microscopy images. My CV-based method was assessed on two real-world microscopy image sets without systematic pre-selection of FOVs, so I am confident that the results comprehensively describe its performance and limitations.

More in-depth analysis reveals valuable insights and contributes to further interpretation of my data. A narrow standard deviation around the mean value of Δ (see Equation 4.5) for both image sets indicates consistency of bacterial shape. However, Image Set 2 exhibits a higher degree of shape variability. This pattern of variation in bacillary shapes raises the possibility that the same method could exhibit strong performance in some aspects of bacillary detection in some datasets but weaker performance in others. For instance, it is a well known fact that the appearances of TB microscopy slides can be highly variable, and even experienced manual microscopists may find it challenging to objectively assess certain FOVs; the effectiveness of the same approach may not be consistent across all instances. Similarly, the results of automated image analysis in this chapter suggest that localisation of bacteria in Image Test Set 1 might have been easier than in Test Set 2 (higher Jaccard Index and SD coefficient) but that shape characterisation might have been easier in Test Set 2 (lower L_1 and L_∞ norm). As previously described in Section 3.3.1, some of the qualitative heterogeneity in slide and microscopy image appearance might be minimised by consistency of laboratory procedures but some of the complexity is driven by inherently complex shapes of *Mtb* cells which cannot be easily controlled.

Although development of geometry-based CV approaches to bacterial detection could begin without annotated training data, use of manual selection criteria for of *Mtb* detection criteria ultimately highlighted a key limitation compared to DL approaches. With DL, model parameters are automatically fine-tuned through data-driven optimisation to best fit the characteristics of the dataset. In contrast, manual tuning relies on human evaluation to handpick parameter values, like the bacteria shape criteria in Table 4.2. This process is limited by the restricted ability of the human mind to comprehensively analyse the full dataset at once to deduce optimal data-specific parameters. DL models can learn a wider range of specialised parameters more efficiently by mining insights across the entire training data during the learning phase. While manual tuning was sufficient for reasonable detection, the lack of fine-grained parameter optimization likely constrained the performance compared to a DL approach. This underscores the potential benefits of incorporating DL for specialized parameter tuning and this approach will be pursued in Chapter 5.

Nevertheless, an important contribution of this chapter to the overall objective of the thesis is that a method of image enhancement was developed which can also be applied to DL techniques. While the utilisation of the ratio of the two eigenvalues may not produce outcomes as favorable as those observed in similar *Mtb* detection approaches, both eigenvalues can be employed to create an enhanced image. This approach serves to reduce noise and effectively smooth the background, thereby accentuating potential objects (bacteria) within the image. More specifically, to create an enhanced image, each pixel in the original image is replaced with the absolute value of the lower-magnitude value of the Hessian eigenvalue computed at the locus. Additionally, it does not affect the morphology of bacteria as the eigenvalue of higher magnitude one does as seen in Figure 4.3. The image enhancement technique developed in this chapter is advanced as a pre-processing step on all images used for training and testing of TB-AI microscopy methods later in the thesis.

Chapter Five

Tuberculosis Bacteria Detection and Counting in Fluorescence Microscopy Images Using a Multi-Stage Deep Learning Pipeline

Chapter Abstract – This chapter utilises the image enhancement technique from Chapter 4 and introduces an autonomous pipeline that uses a new DL-based technique to swiftly detect *Mtb* organisms in sputum samples and estimate the bacillary load. The input of fluorescence microscopy FOVs into a series of networks produces a final count of present bacteria more rapidly and consistently than manual analysis by healthcare. The pipeline consists of four steps: annotation using Cycle-GANs, extraction of salient image patches, classification of the extracted patches, and regression to determine the final bacteria count. Each step of the pipeline is assessed empirically, and a unified assessment is performed using previously unseen data that were labelled with ground-truth values by a microscopist. It is shown that the pipeline can produce the bacterial count for an image sample with an inaccuracy of less than 5% without human involvement.

5.1 Chapter introduction

Chapters 2 and 3 have articulated the potential benefits of, and prior research towards, automation of fluorescence microscopy for clinical monitoring and academic study of TB treatment response. Presently, lack of satisfactory tools for these tasks impairs timely TB diagnosis and delays identification of patients who are at risk of treatment failure [194]. Declining sputum bacterial load once therapy is underway is the most reliable indicator of antibiotic efficacy in pulmonary TB, but mycobacterial culture is difficult to standardise for this purpose and not all viable *Mtb* bacilli are identified [78]. New molecular microbiology assays for rapid TB diagnosis (e.g. Xpert$^{®}$ MTB/RIF) are based on identification of mycobacterial DNA which degrades very slowly and is an unsatisfactory biomarker for monitoring the elimination rate of *Mtb* cells. Efforts are ongoing to develop better molecular tests for TB treatment monitoring but these are not yet ready for widespread use [78].

Section 2.6.1.2 described semi quantitative grading schemes which are used to report the sputum bacterial load. Sputum smear microscopy findings can be reported as “negative”, “scanty”, “1+”, “2+”, or “3+” [128, 95]. This can be helpful for pre-treatment prognostic assessment, as patients with higher bacterial are at greater risk of unfavourable outcomes [108]. Once treatment has started, changes in bacillary load may be noticed as early as 3 days [78]. Therefore, laboratory microscopists may wish to count individual *Mtb* bacteria on sputum smears of samples taken at the time of diagnosis and at regular intervals throughout treatment. Counting *Mtb* bacteria is subjective and time-consuming, exactly like the process of smear microscopy for TB diagnosis. Bacteria are viewed in isolation and usually easy to count, however they often aggregate and create a wide range of appearances. At that moment, the microscopist must make a subjective judgment on whether bacteria are present and how many there are (e.g. whether there are two bacteria with a crescent form or two straight bacteria with a smaller one jammed beside them).

Several of the obstacles associated with manual quantification of bacillary load on microscopy slides may be mitigated by AI tools. Chapter 4 sought to do this using a purely CV and geometry-based approach. This was unsuccessful, but the work unearthed a technique for image enhancement and standardisation. Chapter 5 will combine this image enhancement technique with DL tools to create a hybrid method for *Mtb* classification and quantification. Specific aims of this chapter are to:

- Implement a rapid method for extracting a new representation of microscopic slides, which enhances the differentiation of bacteria from their background.
- Describe a novel method for the detection of salient regions (those which contain *Mtb* cells) within microscopy images, which uses cycle-consistent generative adversarial networks

(Cycle-GANs) to create FOVs with bounding box annotations.

- Introduce a transfer learning-trained convolutional neural network-based refinement of the bacillary detection procedure from the previous step.

- Propose a CNN based method for counting bacteria, including those with variable and atypical appearance, in image patches, using regression as a means of increasing the robustness of the count.

5.2 Image processing-based enhanced representation extraction

Whilst the results of Chapter 4 did not demonstrate sufficient ability of geometry-based CV tools to segment, classify or quantify FOV cells in fluorescence microscopy FOVs, the work described in that chapter was an important step in developing an approach to refine and standardise unprocessed images datasets. Previously enticing CV techniques that take into consideration both eigenvalues of a Hessian matrix, such as the use of the λ_2/λ_1 ratio of the two eigenvalues, were unsatisfactory for bacillary localisation due to the extremely wide dynamic range and the noise that they introduced to image interpretation. However, as described in Chapter 4, when each pixel in the original image is replaced with the absolute lower magnitude Hessian eigenvalue (λ_2) calculated at the locus, an enhanced image is generated. Before employing and evaluating conventional ML and DL tools for bacterial detection, this technique is applied to every image for all experimental work in the remainder of this thesis.

5.3 Generative adversarial networks: A brief introduction

Generative adversarial networks (GANs) constitute a semi-supervised approach to generative modeling that leverages DL algorithms. Specifically, GANs consist of two CNNs: the first undergoes unsupervised training to generate synthetic data, while the second receives supervised training to distinguish original data from generated one. In their default configuration they comprise of two MLP models that are simultaneously trained until convergence. The models are referred to as the Generator (G) and the Discriminator (D). The Generator extracts a random noise vector as input and aims to generate samples that are indistinguishable from real data [248, 89]. The objective of the D is to aid in the learning of G by attentively examining data provided by G [248]. Therefore, it is basically the discriminator's responsibility to determine if the provided data is derived from real data or synthesized by G [248, 89]. The two models are competitors in a *MinMax* game with antagonistic goals who are attempting to fool one another. One of the main advantages of employing GANs for this kind of cross-domain translation include the absence of need for Markov style sampling [143]. Another attractive feature of GANs is the absence of a heuristic function (such as pixel autonomous mean-square error) for representation learning [182]. As an MLP, the model uses back propagation to calculate gradients without requiring any assumptions during learning [89].

Formally, in a standard GANs, G generates a distribution known as the latent space, denoted by $G(z)$. While a Gaussian noise distribution is commonly used, alternative multivariate distributions may also be considered as candidates [129]. Latent space is a random distribution sample in hidden space that groups data points that are closer together [120]. The latent space in a GAN can be conceptualized as a mathematical domain where the generator model synthesizes new data instances, adhering to the statistical properties it has gleaned from the training data. The formulation of the two models is $G(z,\theta_G)$, where θ_G provides the generator's parameters. The discriminator, on the other hand, accepts input from either real data or generated images, whose output can be described by $D(G(z,\theta_G),\theta_D)$, yielding a binary decision as to whether input is from p_{data} or p_G [248, 89]. Model evaluation is based on the following metric:

$$L_{GAN}(D,G) = E_{x \sim p_{data}}[logD(x)] + E_z \sim p_z[log(1-D(G(z)))] \tag{5.1}$$

In the above equation(5.1), as described in the literature, the *MinMax* is formally shown [248]. The first term evaluates the logarithmic probability that the input comes from real data. Hence, the second term evaluates that the input data comes from the generator (synthesised, latent space). It is important to note that the letter E stands for equilibrium which is a constant solution to differential equations, an optimisation technique for MLPs. $D's$ target is to maximise L_{GAN}, naturally evaluating $D(x) \longrightarrow 1$ and $D(G(z)) \longrightarrow 0$. The adversarial objective of G is to minimise L_{GAN} by evaluating the second term to $D(G(z))$:

$$\underset{D}{min}\ \underset{G}{max}\ L_{GAN}(D,G) \tag{5.2}$$

In practise, the *MinMax* game will often fail to get the model to equilibrium, as shown in equation 5.2 [248, 89]. Indeed $log(1-D(G(z))$ swiftly diminishes because of poor quality outputs from early learning of G. Certainly, one does not have any control over what modes of data are being created in a typical GANs frequently leading to unstable output as well [89, 182]. They are also constrained in that the combined architecture of D and G must be fully differentiable, requiring non-discrete values [100]. Therefore, in a default GANs, modifications can be implemented to refine the evaluation criteria for G, focusing on the quality of generated data rather than solely on its ability to deceive D. Changing this aim leads in the suggested separation and expansion of equation 5.2 with the objective of G remaining the same:

$$\max_{G} L_{GAN}(G) \tag{5.3}$$

while the objective of D is also to maximise its performance:

$$\max_{D} L_{GAN}(D,G) \tag{5.4}$$

5.3.1 Transforming GANs to Cycle-GANs

The methodology is proposed by Zhu *et al.* [273]. As implied by its name, the Cycle-GAN variant exhibits cyclical behaviour. To do this, two Generators and two Discriminators are used. Since generators will attempt to map each other's domain, we need two domains. The two domains responsible for this task must be distinct from each other; in this work I define a Labelled domain and an Unlabelled domain (more will be explained later). As described in the original paper [273], assume we have a Generator G that translates images from the Labelled domain (L) to the Unlabelled domain (U). Consequently, Generator G attempts to generate realistic images for the U domain utilising input images from the L domain. The same holds true for the second Generator, F, i.e. from L domain to the U domain. This behavior is modelled as:

$$\begin{aligned} G &: L \mapsto U \\ F &: U \mapsto L \end{aligned} \tag{5.5}$$

As described in section 5.3, the two Discriminators are tasked with determining whether given images are synthetic or genuine, as is the case in default GANs. Each Discriminator evaluates identical images from the same domain. Both Generators and Discriminators are trained using the adversarial loss described in the preceding equation 5.1. Both adversarial loss equations are as follows:

$$\begin{gathered} L_{GAN}(G,D_U,L,U) = E_{u\sim p_{data}(u)}[logD_U(u)] \\ \qquad + E_{l\sim p_{data}(l)}[log(1-D(G(l)))] \\ \\ L_{GAN}(F,D_L,U,L) = E_{l\sim p_{data}(l)}[logD_L(l)] \\ \qquad + E_{u\sim p_{data}(u)}[log(1-D(G(u)))] \end{gathered} \tag{5.6}$$

Both adversarial loss equations, as shown in Equation 5.6, are adaptations from Zhu *et al.,* modified to represent the domains specific to this work, as opposed to their original representation [273]. So far, it seems that two default GANs are being trained with two unique domains simultaneously.

While this may appear to be the case, the unique feature of a Cycle-GANs is that both Generators are also utilised to "unmap" the output of the other. In fact, this is referred to as the cycle-consistency loss, which is computed by inverting the outputs of each generator [273]. To accomplish this behaviour, the authors changed each generator to additionally accept the output of the other as input (as illustrated in 5.5) [273]. Formal description of the retrograde nature is as follows:

$$\begin{aligned} G &: F(L) \approx U \\ F &: G(U) \approx L \end{aligned} \tag{5.7}$$

As mentioned, cycle consistency loss is another form of regularisation to prevent Generators mapping the same input to multiple outputs of the target domain. The full cycle of Generators (forward 5.5 and backward 5.7) is given by $l \mapsto G(l) \mapsto G(F(l)) \approx l$ and $u \mapsto F(u) \mapsto F(G(u)) \approx u$ respectively[273]. Combining the two backward behaviours into one loss is formally expressed as:

$$\begin{aligned} L_{Cycle}(G,F) = E_{l\sim p_{data}(l)}[\ (F(G(l)) - l)^2\ ] \\ + E_{u\sim p_{data}(u)}[\ (G(F(u)) - u)^2\ ] \end{aligned} \tag{5.8}$$

While equivalent to adversarial loss in that they both aim to achieve equilibrium, the means by which they do so varies. In this instance, the concern lies in how well the inverse output matches the real output. Therefore, the primary purpose of cycle loss is to rebuild images as closely as possible to the originals [273].

There is one more loss to consider, although it is supplementary rather than crucial to the overall Cycle-GANs loss. The result is a regularisation of the colour identity mappings of the synthesised images [273]. Without this loss, as shown by the findings of the same paper, the network's behaviour regarding the colour of the synthesised images is unregulated. Identity loss is defined formally as:

$$\begin{aligned} L_{Identity}(G,F) = E_{u\sim p_{data}(u)}[\ \|G(u) - u\|_1\ ] \\ + E_{l\sim p_{data}(l)}[\ \|F(l) - l\|_1\ ] \end{aligned} \tag{5.9}$$

Although both L_{Cycle} and $L_{Identity}$ compute pixel-to-pixel distance error, the former utilises the L2 loss function (MSE) whilst the latter employs the L1 loss function (MAE). The rationale for this modification is that it functions more consistently throughout training and produces higher quality results [273]. Lastly, the overall loss of Cycle-GANs is formally given as:

$$\begin{aligned} L(G,F,D_L,D_U) = L_{GAN}(G,D_U,L,U) \\ + L_{GAN}(F,D_L,U,L) \\ + \lambda L_{Cycle}(G,F) \\ + L_{Identity}(G,F) \end{aligned} \tag{5.10}$$

where the hyperparameter λ controls the relative importance of the cycle consistency loss and its value is determined through empirical testing and is task-dependent. In both this chapter and the original work by Zhu *et al.,* the hyperparameter λ is assigned the value of 10 [273].

5.4 Proposed method

This section explains the key steps of the proposed algorithm in detail, namely i) semantic segmentation of the slide, ii) salient image patch extraction, and iii) regression-based inference of the bacterial count from the extracted patches.

5.4.1 Object detection using Cycle-GANs

The primary factor for the development of Cycle-GANs was its capability for image-to-image translation. In contrast, the objective of this application is to train Cycle-GANs to transfer labels from labelled (L domain) images to unlabelled ones (U domain). The L domain, which consists of microscopic fields specifically labelled by an expert with regions of interest, and the U domain, which consists of microscopic fields without labels. Certainly, the opposite behaviour will also occur as well, but this is of no importance for this context of work. Additionally, the Cycle-GANs cyclical nature is maintained largely for regularisation and improved overall performance. Ultimately, synthetically created labelled images are utilised to assess the entire performance.

Following experimental results reported in previous work [273], my method initially used input image patches with the size of 256×256 pixels but additionally re-scale them to 384×384 pixels using bicubic interpolation [8], which was found to effect an improvement in performance. I also introduce alterations to the network architecture by including three further residual blocks as a means of improving the detection of bacteria with lower brightness. This larger input size can also reduce information loss during the encoding and decoding processes, while generating more robust and spatially rich feature maps, albeit at a higher computational cost. It also improves synthetic image quality by capturing finer-grained features and expanding the receptive field, thereby providing a more comprehensive context for each pixel.

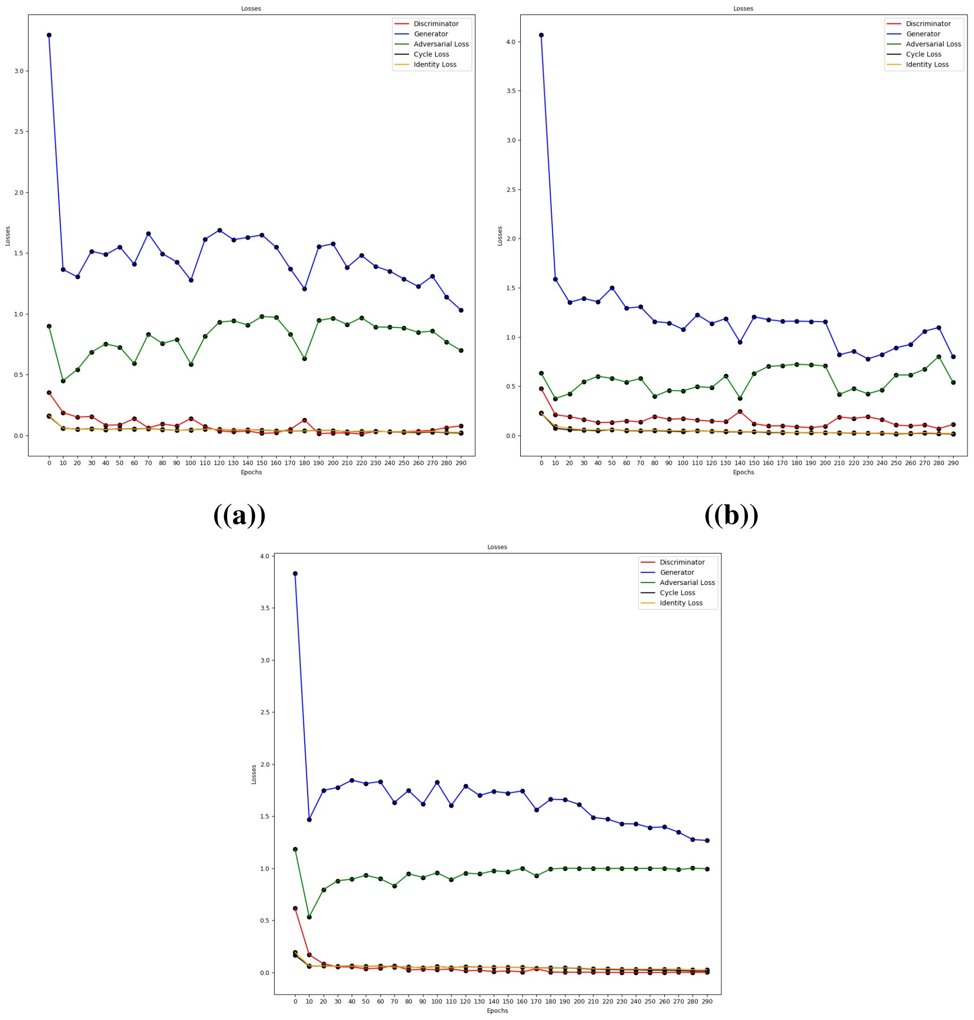
As regards the discriminators, which classify overlapping patches, I adopt an architecture similar to that of the PatchGAN [109, 123, 125]. However, evidence shows that the relatively large patch size (70×70 pixels) used by most previous work is unsuitable for the context of tasks in which the generators are trying characteristics that are more granular and nuanced in appearance [267]. Hence, I use much smaller 30×30 pixel patches herein instead. Additionally, to further increase the sensitivity and robustness of the model, I introduce a change to the usual number of strides at different layers. In particular, as a means of facilitating the learning in the proximity of the image border, I introduce a reflection pad of size 3. Table 5.1 summarizes the key changes.

Table 5.1: Key parameters of the five-layer discriminators used in the present work. Changes from the usual values used in previous work are shown without highlighting, whereas this task-specific alterations are shown using bold font. In this task, the transition from 2×2 to 3×3 kernels was motivated by the necessity for finer-grained feature extraction. Bacteria, as the objects of interest, require a more detailed spatial understanding, and the 3×3 kernels with a centered pixel offer a more suitable framework for capturing intricate patterns and details in the images.

Layer	Kernel Size	Strides	Padding
Layer 1	$\mathbf{3\times3}$	2	**3**
Layer 2	$\mathbf{3\times3}$	**1**	1
Layer 3	$\mathbf{3\times3}$	**1**	1
Layer 4	$\mathbf{3\times3}$	**3**	1
Layer 5	$\mathbf{3\times3}$	**2**	1

5.4.1.1 Training the Cycle-GANs

Here, I summarise the key settings pertaining to the training of the cycle-GAN. To start with, considering the complexity of the learning task at hand, the number of epochs used to train the cycle-GAN was set to 300, which is a considerably higher number than that used in most previous work [273]. Another crucial aspect of training which needs to be correctly determined to facilitate successful cycle-GAN learning concerns the learning rates of the generators and discriminators. In particular, a sense of competitive equilibrium has to be maintained between the two kinds of sub-network. If the discriminators are considerably more effective, the network will overfit and the generators' learning will never converge. Similarly, if generators are more effective, mode collapse is likely, and the desired state of the overall network may never be achieved. Other works that employ cycle-GANs for highly specialised tasks have shown the benefit of differing learning rates for the two sub-networks [267, 79]. Similarly, the learning rate of the generators was set to 0.0006 and that of the discriminators to 0.0002. These considerations led me to effect a reduction in the (linear) learning rate after 50 epochs—significantly earlier than most other work [273]. I also adopt the use of AdaBelief, a new optimizer which has shown to converge as quickly as adaptive optimizers (such as Adam [117]) and to generalize better than Stochastic Gradient Descent (SGD) [187] in complex architectures such as GANs [274]; see Figure 5.1.



((a))

((b))

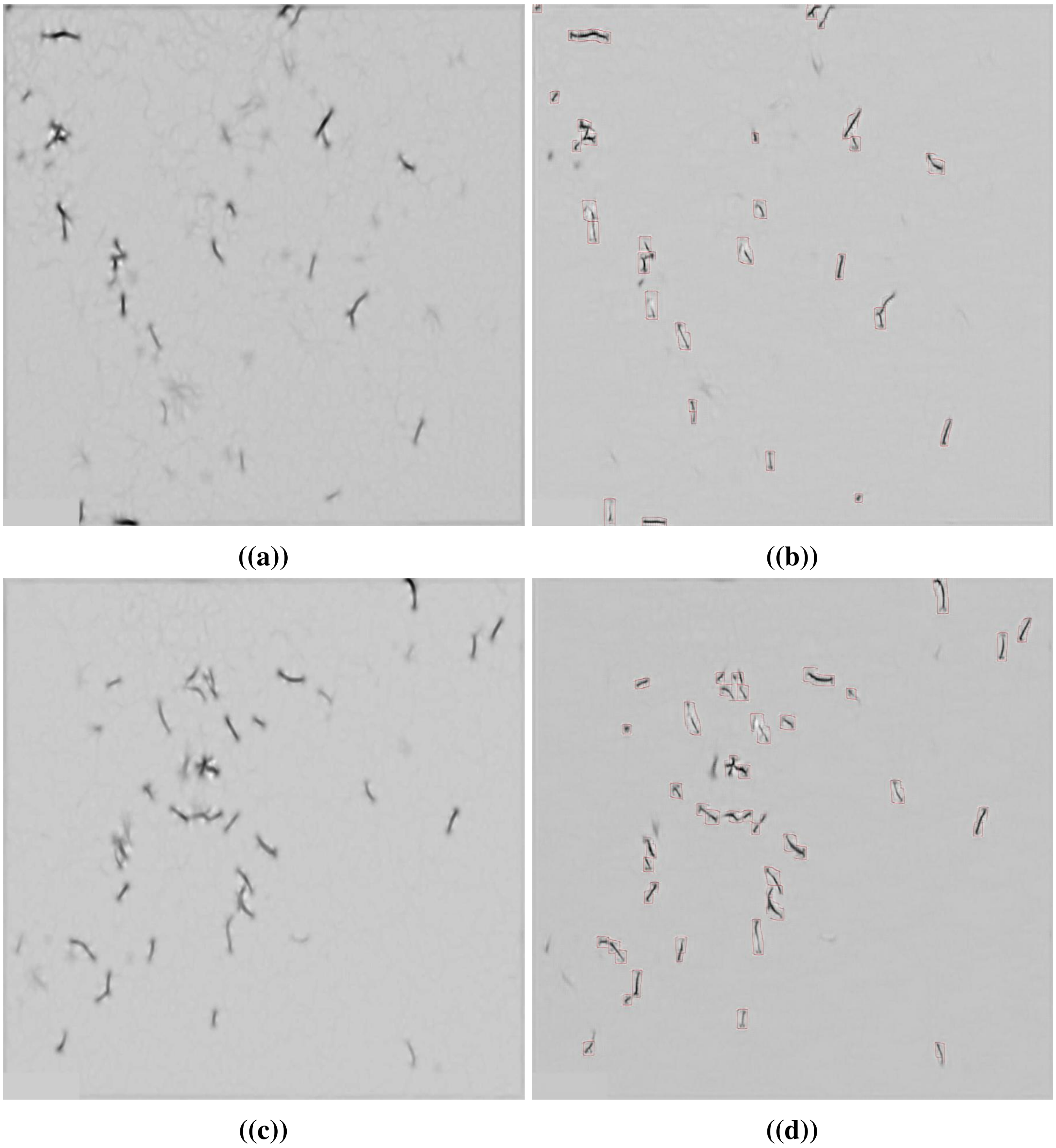
((c))

Figure 5.1: Comparison of training losses observed with the use of different optimizers. Note that the adopted AdaBelief effects the smoothest learning behaviour. (**a**) Adam; (**b**) SGD; (**c**) AdaBelief.

Finally, to maximize the robustness and the generalizability of the learning process, I perform synthetic data augmentation. In particular, I increase the amount of training data by approximately 50% by adding images randomly rotated by ±25° and reflected about the vertical or the horizontal axis [260]. This kind of augmentation is particularly suitable for TB microscopy image analysis because, unlike in the case of natural images wherein there is an inherent asymmetry in directions (e.g., the horizontal and vertical directions are objectively defined and cannot be swapped one for another), in the microscopy slides of interest here, all directions are interchangeable and in that sense equivalent.

5.4.2 Extracting salient patches from synthetically labelled images

Considering that the images based on the enhanced representation described in Section 4.4 are greyscale and the superimposed bounding box red, the localization of the former is a rather straightforward task; see Figure 5.2. I start by simple colour thresholding, localizing pixels with the red channel value between 150 and 215 (within the range of 0–255), and the green and blue channel values between 90 and 160. The subsequent application of morphological dilation and erosion ensures that the extracted salient structures, which correspond to bounding box contours, are properly closed, thus suppressing the effects of noise.

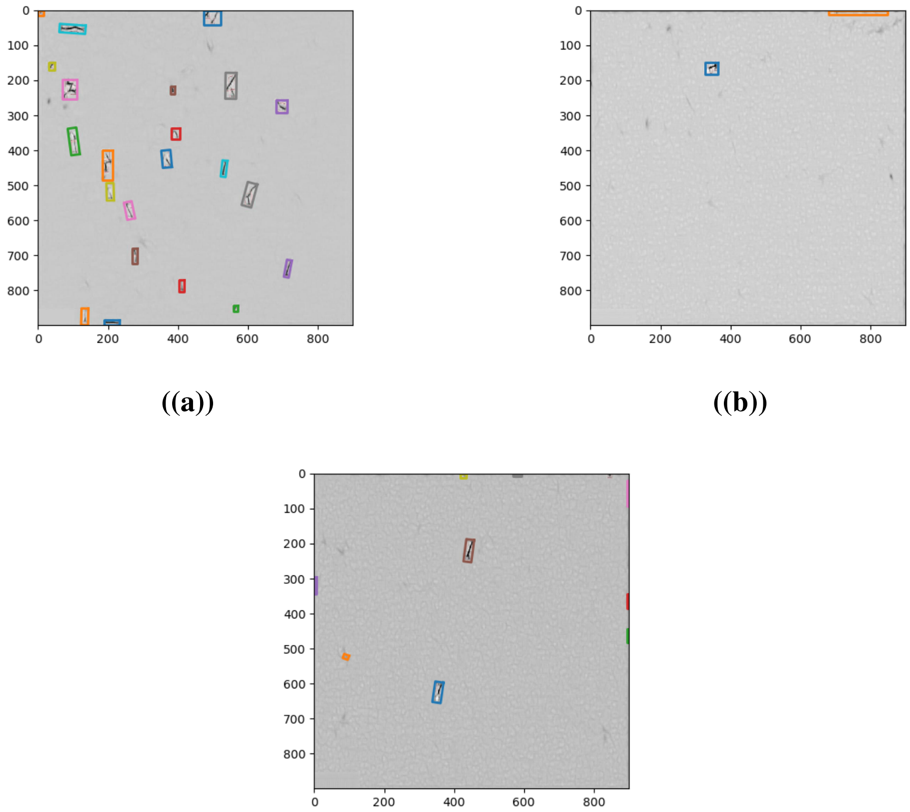


((a)) **((b))**

((c)) **((d))**

Figure 5.2: Examples of (**a**,**c**) complex and cluttered original input images and (**b**,**d**) the corresponding output images generated using the proposed cycle-GAN, showing synthetically superimposed bounding boxes around the bacterial content of interest.

To further increase the robustness of my approach, I follow the aforementioned low-level processing with a more semantic, domain knowledge-driven refinement. More specifically, guided by the understanding of *Mtb* cell dimensions in slides, I impose certain constraints on the extracted bounding boxes. Using the Douglas–Peucker algorithm [67], a polygonal approximation of imperfectly extracted and possibly overlapping bounded boxes is computed, and any candidate object with a perimeter outside the range of 60–450 pixels is rejected. Finally, I extract patches of interest using minimal bounding boxes enveloping the convex hulls of all connected salient structures; see Figure 5.3.



((a))

((b))

((c))

Figure 5.3: Examples of (**a**–**c**) with minimal bounding boxes drawn based on the criteria described. Note that certain structures generated from noise continue to incorrectly meet the aforementioned requirements, often near the image boundary evidenced by (b, c). This further influences the need for an intermediate classification stage.

5.4.3 Classifying cropped patches

Until this point, the aim of the method is to extract as many patches that were of sufficiently bacteria-like objects by using a rather coarse criterion that facilitates fast processing. For this reason, the acceptance level is set favouring sensitivity to capture all FOV, which means that some false positives will also be selected. The goal of the next phase is then to determine whether a selected bacterium patch is a true positive by using more nuanced local appearance. This is challenging because of varied atypical and overlapping bacillary appearances, as described in Chapter 4. In order to address this variability, I pursued an DL approach whereby the discrimination between bacterial and non-bacterial patches was formulated as a classification problem, which was solved using a CNN.

To this end, I apply and compare several of state-of-the-art models, namely the ResNet family [98], the DenseNet family [104], and the SqueezeNet1_1 family [107]. Each model's first convolutional layer is replaced with one that consisted of one input channel, kernel 3×3, stride 1, and three 3×3 padding. The alterations are motivated by the fact that the slide representation was monochrome (that is, single channel) and the objects of interest are thin, elongated structures that frequently appeared near the image boundary. Every model's last linear layer is replaced with a single-output linear layer. The linear layer's output weights are then fed into the sigmoid function. Finally, binary cross entropy is employed as the loss function, and the models are pre-trained on ImageNet.

Precisely 5194 patch images are used for training, with a balanced split between positive and negative examples. Positive examples are extracted using the method explained in Section 5.4.2, whereas negative ones are selected by randomly sampling from the FOVs and accepting those patches which did not overlap with any of the positive ones. Specifically 770 images are used for validation. A three-pixel-wide frame is constructed on a randomly chosen positive image (which was known to contain bacteria) to approximate the boundary box formed from the projected labels and to prevent overfitting on the training data. The learning rate is set to 0.0001, with a circular scheduler that had a step size equal to five times the size of the dataset (which in turn is dependent on the batch size) [208]. The base learning rate and the upper learning rate are set to 0.0001 and 0.0002, respectively. Stochastic gradient descent is used as the optimizer since it has been demonstrated to generalize better than Adam [117] in related image classification problems [274]. The model is trained for 100 epochs, with a 0.03 loss and accuracy tolerance, resulting in the termination of training following 20 epochs of no improvement.

5.4.4 Counting bacteria

In the final stage of my algorithm, I use regression to infer the number of bacteria present in an input image patch. As I will explain in more detail in the next section, I compare a number of different architectures and modify them all by replacing their last linear layer with a single output layer. The MSE loss function is used for training, and Adam [45, 274] is used as the optimizer, with a circular scheduler having the lower and upper boundaries of 0.0001 and 0.00015, respectively; the step size used is equal to twice the size of the dataset. Because patches with more than three bacteria are exceedingly uncommon, I use a relatively low batch size that resulted in a model update following every few examples, thus avoiding the dominance of patches containing a single or two bacteria. Therefore, the batch size is set to 22, or about 5% of the dataset size, in order to maximize the generalizability of the learning.

5.5 Experimental evaluation

In this section, I describe an empirical assessment of the proposed algorithm using real-world data. I begin with a description of the data used and follow up with an ablation study of the different stages of the pipeline.

5.5.1 Dataset

For all experiments in this chapter, Auramine O stained FOVs from Image Test Set 1 (described in Section 4.3.2) were used.

500 FOVs were selected across all time points of sample collection to ensure that the automated detection and counting networks for FOV bacteria presented here would not be confounded by any changes in bacillary morphology during TB treatment. These images were reviewed within an annotation tool for image labelling by a microscopist who had not previously seen these data and was not involved in development of the DL algorithm. Rectangular bounding boxes were superimposed around bacteria within each image to tag areas of interest which contained one or more FOV cells as well as the number of cells present. Overlapping boxes were merged; see Figure 5.4. All FOVs were enhanced using the technique mentioned in Sections 4.4 and 5.2

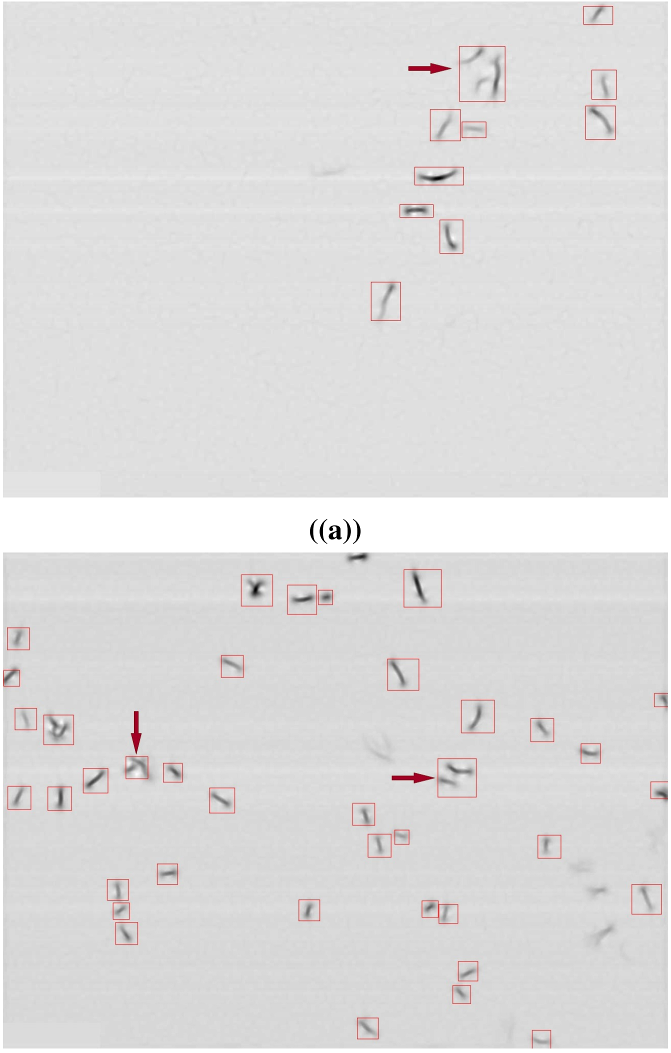


Figure 5.4: Examples (a, b) of bounding boxes created by a trained microscopist around areas of interest. Red arrows illustrate the union of two boxes that overlap.

5.5.2 Results

To facilitate an in-depth, nuanced understanding of each stage in the proposed pipeline I performed an ablation study, that is I evaluated each stage of my algorithm in turn and discussed its contribution to the overall performance [154].

5.5.2.1 Semantic segmentation using Cycle-GANs

To gain insight into the performance of the semantic segmentation, I examined the overlap between ground truth segmentation and that achieved using my automatic method. In other words, I was interested in quantifying the degree of coincidence between two binary images, each comprising regions of interest and the remaining image content, as illustrated in Figure 5.5.

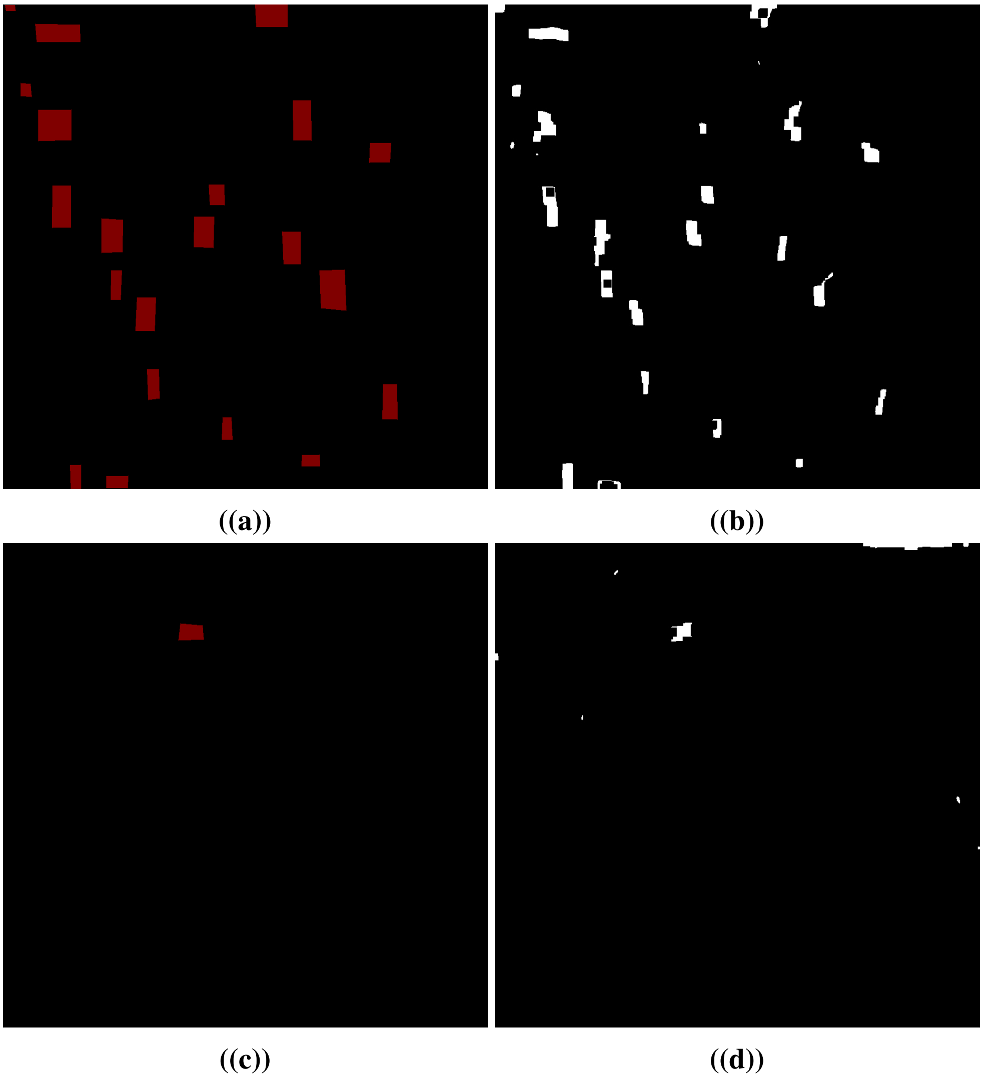


Figure 5.5: Examples of (**a, c**) ground truth and (**b, d**) the corresponding predicted salient FOV regions containing possible MTB cells shown as binary images.

I started by looking at the standard metrics for this kind of assessment, namely the Jaccard index [21] and the SD coefficient [93]. On the test set, I found these to be 89% and 94%, respectively, suggesting highly effective performance. Considering that, in this thesis Cycle-GANs is the exclusive model able to process RGB images due to its three-channel input configuration (intended to reproduce red bounding boxes even though the enhanced images are greyscale), it was suitable for empirical scrutiny to ascertain whether the image enhancement technique indeed increases the detection of *Mtb* bacteria. Hence, I conducted training for Cycle-GANs using Auramine O FOVs while keeping the settings identical to those used for training on the enhanced images to facilitate fair comparison. In this case the results were 87% and 93%. Overall, after examining the data manually, I found that the deviation from perfect performance was due to boundary effects, which is the slight misalignment of the exact boundaries between the ground truth and the predicted regions of interest rather than an entirely mistaken focus. Moreover, Cycle-GANs sometimes has a tendency to inaccurately translate bacteria-like objects in the generated images (along with bounding boxes), particularly in relation to the contrast of the the background; see Appendix B.1 for more details.

To test the observation, I next introduced a custom performance metric, designed specifically for the task at hand. In particular, I devised a way of deeming each detected salient region as correct or not, allowing me to quantify the number of false positive and false negative patches, as well as the distance (error) between each true positive and the corresponding ground truth. To do this, I computed the centroid of each predicted salient region, and if possible, coupled it with the centroid of a ground truth salient region. To determine the pairing, the Euclidean distance between each predicted centroid and all ground truth centroids was calculated, and the nearest one was selected as the correct one. A distance threshold of 5 pixels was also used to reject the coupling of centroids that were excessively far apart. Unpaired predicted regions were considered as false positives in the model prediction. Similarly, unpaired ground truth centroids were considered as false negatives in the model prediction.

Out of 294 ground truth centroids, 3 were not paired, and out of 331 labelled predictions, 40 were not paired. The L_1, L_2, and L_∞ distances between paired centroids were found to be 310.49, 18.82, and 1.89 pixels, respectively. As the typical length of a single bacillus ranges from 40 to 140 pixels, these numbers corroborated the previous observation that the segmentation was successful, and that the errors suggested by the Jaccard index were mostly due to small misalignments between the predicted and ground truth salient regions. Such errors had little effect on the performance of the entire pipeline as they did not change the actual bacterial count in the patches passed for further processing.

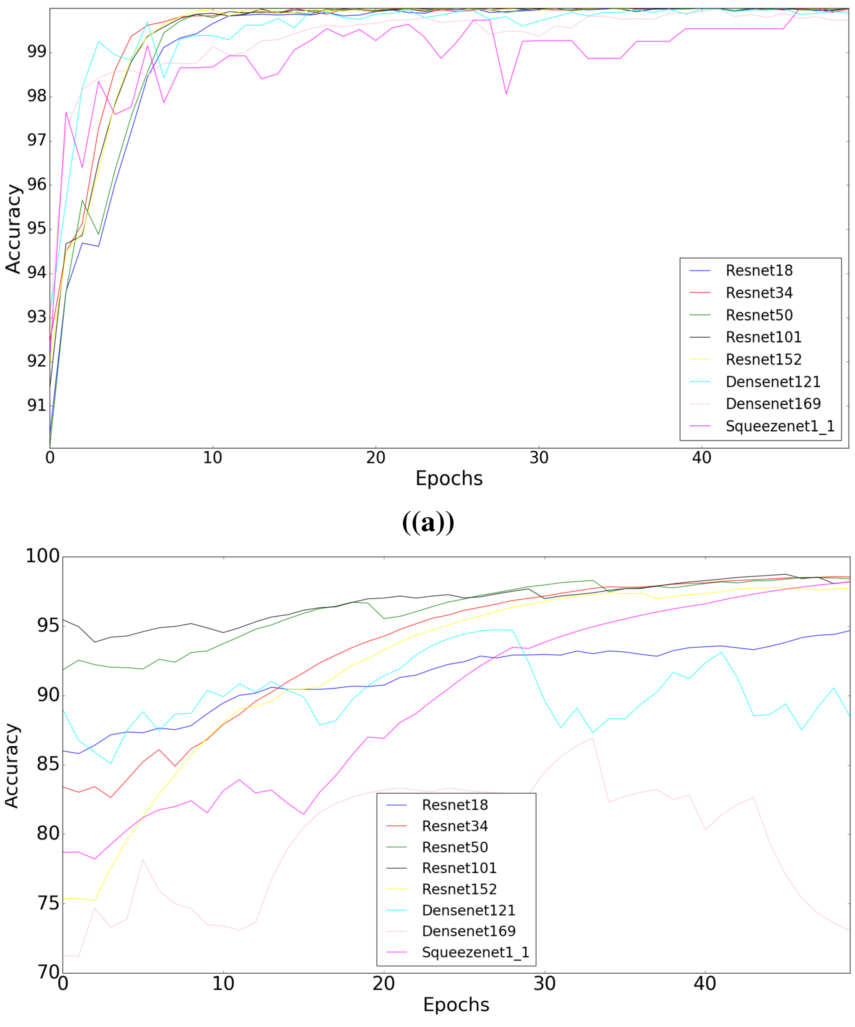
5.5.2.2 Deep learning-based patch classification

Turning attention to the analysis of the second stage of my algorithm, I assessed the more nuanced, DL-based classification of candidate patches as bacteria-containing candidate patches and those void of bacterial content. Using the baseline model, I compared a wide range of different architectures, namely ResNet [98], DenseNet [104], and SqueezNet [107] using standard performance evaluation metrics for classification, all modified as per Section 5.4.3. Following training and validation, I evaluated only the model on the test set that performed best during the validation.

Table 5.2: Validation accuracy achieved by different models. Bold font is used to highlight the best performance according to different criteria (columns).

Model	Accuracy	Precision	Recall	F1-Score
ResNet18	97.28%	0.974	0.949	0.961
ResNet34	99.35%	0.970	0.951	0.960
ResNet50	**99.74%**	**0.990**	**0.967**	0.960
ResNet101	99.61%	0.983	0.958	**0.970**
ResNet152	99.48%	0.980	0.954	0.967
DenseNet121	95.20%	0.952	0.928	0.939
DenseNet169	88.41%	0.900	0.849	0.874
SqueezeNet	99.38%	0.980	0.958	0.969

During training, all models reached 100% accuracy; see Figure 5.6. Greater differentiation was observed during validation, with ResNet50 achieving the highest accuracy of 99.74%, see Table 5.2. Other ResNet models also performed well, as did SqueezeNet, with the exception of the shallowest ResNet18. Both DenseNet models were less successful, and interestingly, the deeper DenseNet169 in particular. Overrall deeper models performed worse, with the validation accuracy decreasing together with the network depth.



((a))

((b))

Figure 5.6: (**a**) Training and (**b**) validation accuracy across epochs of the compared models based on different modified architectures. Interestingly, deeper models performed worse, with the validation accuracy decreasing together with the network depth.

Having been identified as the best performing model during validation, I henceforth adopted ResNet50 as the classifier to evaluate the test set. In summary, I found that the proposed ML-based filtering increased the overall specificity of the pipeline in the discrimination between bacteria-containing patches and those void of bacteria, from 89% attained at the previous, coarse filtering stage, up to 97%. Similarly, sensitivity was increased to 99%, which, exceeded the performance of previously published works [73, 75, 114, 270, 83, 257].

5.5.2.3 Bacterial counting

The final stage of my algorithmic pipeline concerns the counting of bacteria in the salient patches correctly identified in preceding stages. Recall from Section 3.4.2 that regression analysis can be valuable for this task because it predicts real numbers (which may be fractions), even although the actual count can only possibly be an integer. The decision to apply regression here was motivated by the desire to retain information about the uncertainty involved in inferring the bacterial count. Thus, the predicted pseudo-count of 1.05 can be interpreted as more confidently corresponding to a single bacterium than, say, 1.48 (whereas 1.51 would tilt the decision towards the count of 2). My approach also allows for the cancellation of uncorrelated errors across the slide, as observed in previous research [241].

A summary of my experimental results is shown in Table 5.3. The best performance was obtained using the simplest and shallowest model, namely ResNet18. Its error of less than 5% is a significant improvement on all previous work, and attains a new state of the art [212, 147, 241]. The visualizations shown in Figure 5.7 provide further insight into the learning achieved using ResNet18. Both the activation maps and the ultimate count predictions confirm that the network is correctly capturing salient content and appropriately utilizing it to form the ultimate prediction.

Interestingly, note that all models in Table 5.3 overestimate the bacterial count (the aforementioned ResNet18 the least so). To understand why this is the case, in addition to the ultimate assessment criterion, which is the accuracy of the final count, I include in the table three additional metrics computed during training, namely the MSE, the MAE, and the coefficient of determination (R^2). Indeed, an examination of the last of these suggests that overly flexible models, which are very deep models with higher numbers of free parameters, overfit during training.

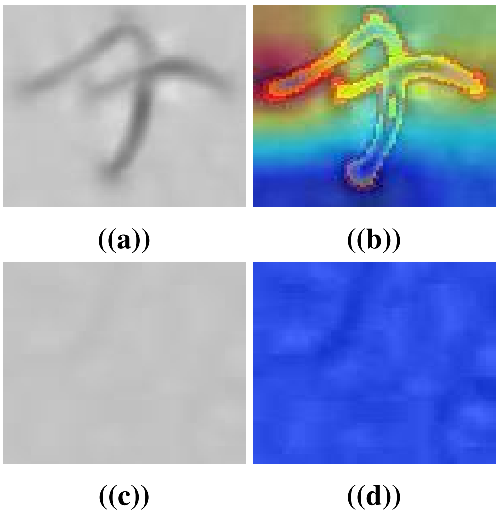


Figure 5.7: GradCAM visualization of trained ResNet18's last layer response to different types of input. Shown are (**a**) an input patch containing three unusually shaped clumped bacteria and (**b**) the corresponding bottleneck layer activations, which show the highest responses around the most salient content; (**c**) a background patch and (**d**) the corresponding bottleneck layer activations, which are nearly non-existent. As expected, the patch in (**a**) results in the regression prediction for the bacterial count of 2.847, and the patch in (**c**) for a count of 0.0256.

Table 5.3: Performance statistics on unseen test data (second column), and training statistics (columns 3–5). Observe that the more flexible, deeper models tend to overfit and thus perform less well on novel data. This is demonstrated by the training R^2 metric, which is low for these models.

	Test	**Training**		
Model	**Count (Ground Truth = 377)**	MSE	**MAE**	Variance lost(1-R^2)
ResNet18	**394**	**0.0054**	0.0345	0.006439
ResNet34	407	0.0444	0.0457	0.006506
ResNet50	414	0.0457	0.0425	0.006523
ResNet101	431	0.0253	0.0236	0.000656
ResNet152	496	0.0231	**0.0201**	**0.000095**
DenseNet121	575	0.0104	0.0603	0.000345
DenseNet169	667	0.0086	0.0406	0.000356
SqueezeNet1_1	404	0.0082	0.0227	0.006571

5.6 Discussion

This chapter has demonstrated how a novel multi-stage DL pipeline approach can detect bacteria with a range of morphologies, unlike previous methods which assumed a much more uniform appearance [75, 114, 212, 147], while also exhibiting greater robustness in challenging conditions, owing to the probabilistic nature of the inference at its crux.

I suggest conducting further experiments to explore the potential of the proposed image enhancement technique. This should involve incorporating additional datasets, including those captured through brightfield microscopy, in order to assess the technique's ability to standardize and enhance the detection of *Mtb* bacteria across different dye types and microscopy methods. Although the first observations from Section 5.5.2.1 do not provide conclusive evidence however they indicate significant visual improvement in several samples. Firstly, the novelty of the proposed approach lies in its demonstration that the color of the FOV has no bearing on the localization of *Mtb* bacteria. Paradoxically, this was not the case in this chapter, as Cycle-GANs were trained on RGB images to retain the red color of the bounding boxes. Secondly, building upon the previous point, it opens the possibility that a single model can accept input FOVs from various dyes or even different microscopy types. Lastly, it represents the initial stride towards the standardization and improvement of FOVs for TB-AI, as it ensures that the same type of image is used by various comparable methods (although numerous challenges remain in dataset standardization). In the context of image improvement, the proposed approach has a more pronounced impact on lower-quality FOVs, i.e. FOVs with greater noise and poorer background contrast, compared to higher-quality ones.

As evidenced by the experiments of the classification and regression stages of the pipeline, the utilization of CNN generic models, particularly those of greater depth, tends to lead to the phenomenon of overfitting in the results. Indeed, I observed a trend where the extent of overfitting decreases as the model's depth is reduced. This phenomenon is reflected in the remarkably low values of the MSE and MAE metrics, as well as the very high value of the R^2. In predictive modeling, encountering a situation where the number of model parameters exceeds the data points and this can increase the risk of overfitting [110]. From a conceptual perspective, each model parameters (or weights), serves as a control point. When the number of control points exceeds the number of observations, the model has the potential to perfectly fit the training dataset[22]. However, this inclination toward overfitting becomes apparent when the model performs exceptionally well on the training data but falters on new, unseen data. Moreover, the observed results within this chapter underscores that the number of features characterizing bacterial detection is considerably fewer than models parameters. This observation underscores the significant role of custom architecture in DL models, a topic that will be elaborated upon in the following chapter.

The findings drawn from this chapter hold promise for both clinical and academic applications of the proposed method. By minimising the need for extensive human involvement, automation of *Mtb* cell classification and quantification can be achieved. This approach can save time and add objectivity to interpretation of fluorescence microscopy results and has potential to be developed further for monitoring of patient response to programmatic and experimental TB therapies, as outlined in Chapter 2. However, a current constraint to the current technique is that it starts with examination of pre-selected fluorescence microscopy FOVs which still require laborious manual microscopy of full slides to assemble. The need to ‘pick’ the FOVs for automated analysis also remains to be vulnerable to the risk of (even unconscious) selection bias in preparing the dataset. Future work, in Chapter 6, will consider how automation of whole microscope slide images could be undertaken to try and manage that problem.

Chapter Six

Extracting and Classifying Salient Fields of View From Microscopy Slides of Tuberculosis Bacteria

Chapter Abstract – After preparing and staining sputum smears, it is a subjective and time-consuming process for laboratory microscopists to view slides and capture FOV images for the types of bacterial detection and quantification analyses described in this thesis. The objective of this chapter is to demonstrate a method which can capture Whole Slide Images (WSIs), then crop salient FOVs from them for further processing. Auramine O LTR stained slides are used for this because of the ultimate combined academic goal to detect *Mtb* cells and describe their lipid content. The method developed uses the same image enhancement tool as earlier chapters, which also converts each FOV to greyscale in order that cell detection models do not need further training to distinguish bacteria stained with different coloured dyes. A bespoke model consisting of two encoders and one classifier is created to detect *Mtb* bacilli. Using data from the Image Set 2, the proposed method is shown to outperform 12 existing methods for FOV classification on the two key metrics, achieving i) an approximately 10% lower overall error rate than the next best model and ii) 100% specificity (with the next best model achieving the specificity of 93%).

6.1 Chapter introduction

Chapter 5 describes a successful approach to automated classification and quantification of *Mtb* cells on sputum smear microscopy FOV images. However, the process of manually creating these FOVs for analysis, in itself, is time-consuming and subjective. Once the sputum smear is fixed and stained (according to the process described in Appendix A.1), a microscopist must scan the whole slide at either ×400 or ×1000 magnification using a green channel filter to identify Auramine O stained bacilli. Paired digital photographs must then be taken (capturing both green and red channel images if intracellular lipid assessment is to be done) of every FOV containing one or more possible bacterial cell. In addition to requiring pain-staking attention to detail (which slows the procedure down to such an extent that the value of AI for image interpretation is reduced), any errors or sub-conscious bias in slide-reading by the microscopist will influence the image-set obtained and compromise all subsequent analyses.

If TB-AI microscopy is to be useful, there is an obvious need to automate the process of FOV selection from microscope slides, decreasing (or eliminating) human involvement and reducing the amount of time required to execute this phase. This problem was previously identified in Section 3.8. Ideally, the first step in slide ‘screening’ should achieve a balance between sensitivity and specificity so that all FOVs containing objects which are ‘potential bacilli’ are retained for more detailed evaluation. Selection of some false positive FOVs is acceptable if it avoids discarding false negative fields (in which bacteria are present but missed) but advancing too many false positive fields through initial screening reduces the value of this procedure in reducing the volume of downstream work. As noted in Section 3.8, some researchers have investigated autofocusing algorithms, for initial slide screening in pulmonary TB diagnosis. In general, this approach has not been successful. The full range of variable morphology and colour intensity of *Mtb* cells under fluorescence microscopy is difficult to capture on autofocus algorithms, so bacteria can be missed.

An alternative approach, if the relevant equipment is available, is to take a whole slide image (WSI) of stained sputum smears which can then be sub-divided into smaller sections for assessment by DL tools. This method has been considered for other WSI AI applications [62], but has rarely been used in TB-AI research. A challenge is that WSI image dimensions are large, ranging from 15-20GB in total size. Considerable computational power and memory is required to handle such images, including training of DL tools to analyse them. Class imbalance training is also a challenge: as previously described, the majority of TB microscopy slides consist of background, which can cause AI classifiers to be skewed towards negative classification and, thus, not generalise effectively, as shown in the work of Kant and Srivastava [114].

The objective of this chapter is to develop:

- A technique that crops WSIs from sputum smears of TB patients into FOV of similar size.
- A filtering step that eliminates non-salient FOVs, retaining only those containing potential *Mtb* cells in the final image dataset.

6.2 Division of WSIs into FOVs

Two digital microscopy WSIs for the work of this chapter came from Image Set 2, as described in Section 4.3.2. To recap, Auramine O LTR stained whole slides were photographed using a Zeiss Axioscan Z1 scanner with Zeiss Zen 3.1 software. Optical magnification of 40 was used, supplemented by digital zoom on the scanner. WSIs were too large ($\approx$ 19GB) for Zeiss Zen 3.1 to handle, so the software divided each slide into 2700 tiles. These were further cropped into 200×400 pixel patches which were anisotropically scaled by a factor of 4.83 in the x direction and 3.24 in the y direction to match the FOV size manually produced by a microscopist in Image Set 1.

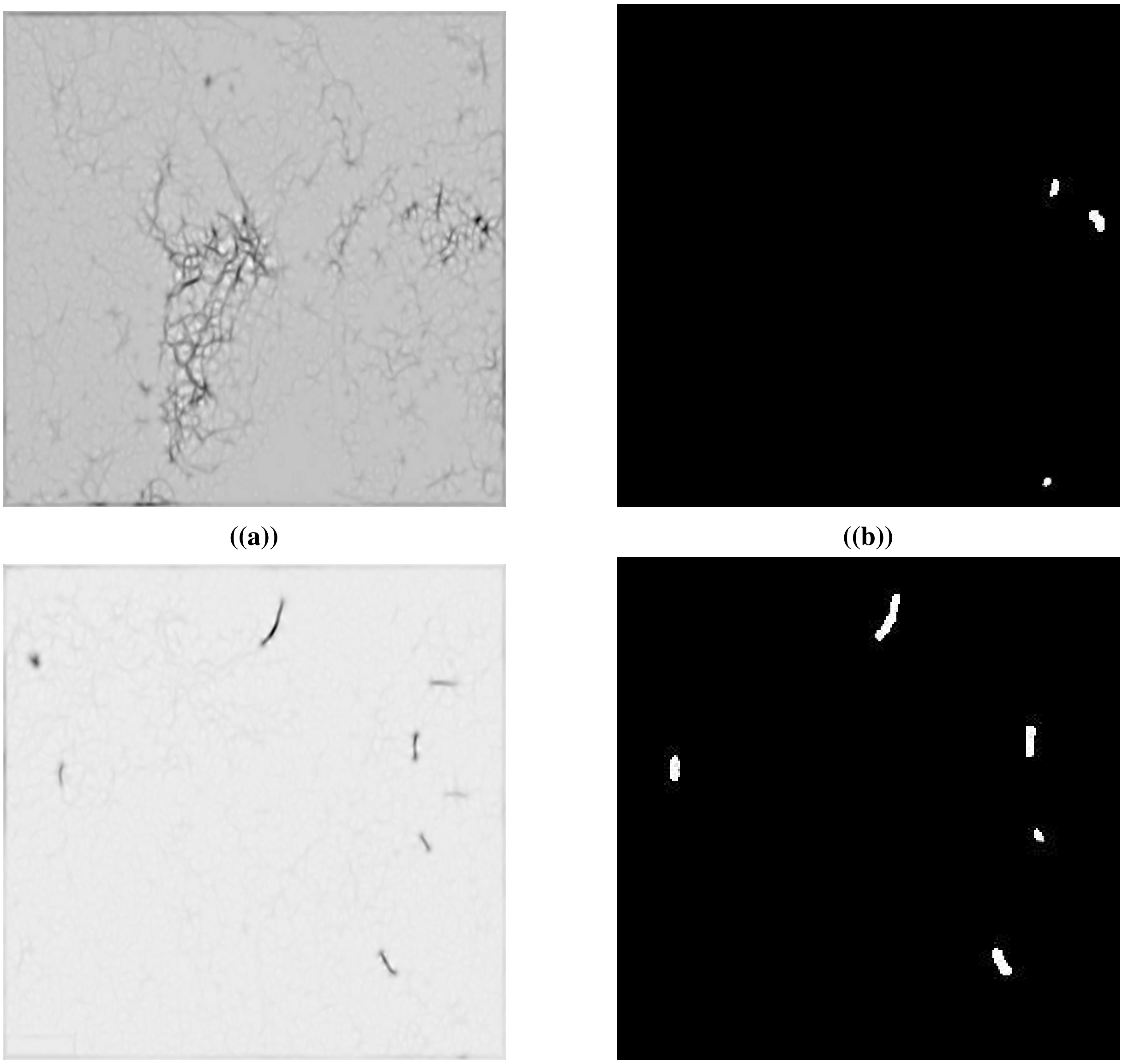
As described above, to obtain paired Auramine O and LTR images, manual microscopy requires physical capture of two separate digital photographs of each FOV viewed through different filter cubes. The process for WSIs via the Axioscanner is slightly different. A single digital whole slide scan is obtained, but two images are produced for each FOV by adjusting RGB colour channels afterwards. Setting the blue and red channels to 0 (by suppressing, not erasing them) leaves information in the green channel intact for Auramine O stained AFB detection. Suppressing the blue and green channels to 0, leaves information in the red channel intact for visualisation of lipids.

6.3 DL for FOV classification

The end goal is to classify cropped FOVs as positive or negative for *Mtb* bacteria. By filtering out FOVs which definitely do not contain objects of interest, the subsequent burden of work required to be done by a microscopist or AI method is reduced.

6.3.1 Proposed model

Evidence from previous work on non-automatic sputum smear microscopy analysis, suggests that for the detection of *Mtb* bacteria the use of both pixel intensity and shape information is superior to the use of either of the two in isolation [242, 75, 212]. Herein I introduce custom-designed networks that reflect this finding by employing two encoders to generate two separate feature maps. One of these is trained on the discrimination enhanced representation of FOVs introduced in Chapter 4 and explained further in Chapter 5, while the other is trained on the binary images (also known as binary masks) corresponding to each FOVs, which distinguish between the objects of interest (bacteria) and the uninteresting content (background and artefacts); see Figure 6.1. The encoder outputs are concatenated to generate the input matrix for another smaller network (16×32×512 pixels); see Figure 6.2. The weights of the two encoders are frozen, and no gradient computation is done during the training of the smaller network. As a result, the smaller network makes an effort to infer the probability distribution from the two encoders which independently infer pixel intensity and shape. To train the two encoders, a further layer with adaptive max pooling and a linear layer leading to a single output unit with a sigmoid activation function are added. The same environmental and hyper-parameters as previously employed are utilized to train the two encoders and the smaller network.



((**a**)) ((**b**))

((**c**)) ((**d**))

Figure 6.1: Examples of (**a**,**c**) FOVs along with their corresponding (**b**, **d**) binary mask counterparts. Even though there are discernible forms in (**b**), when combined with the relevant intensities derived as feature maps from (**a**), the network learns that these shapes are not bacteria.

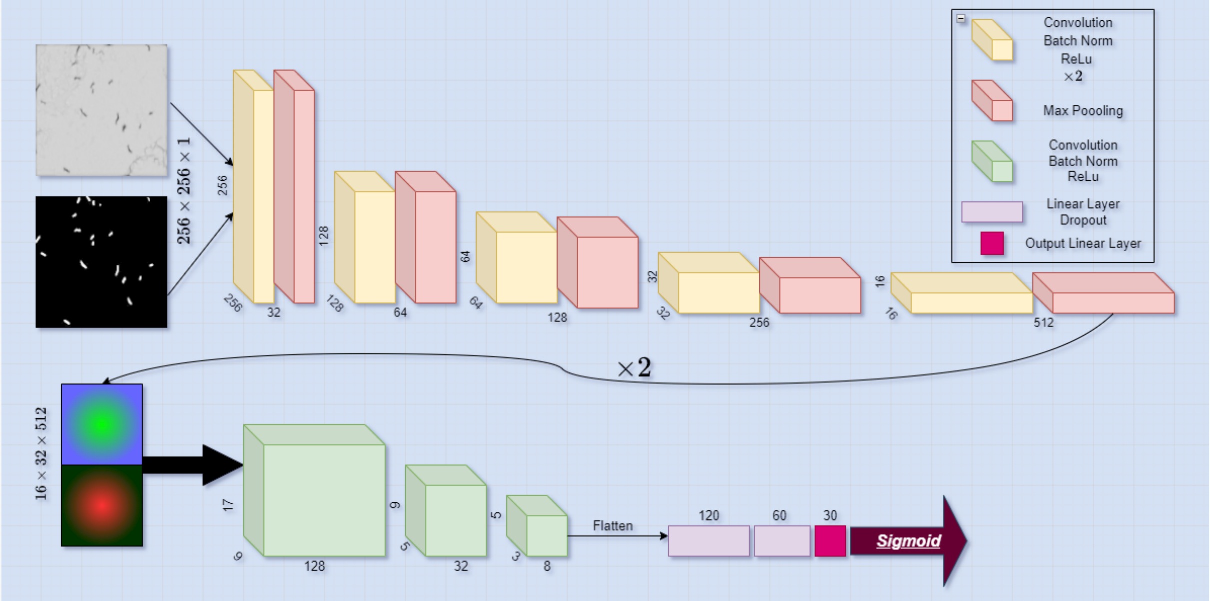


Figure 6.2: Diagram of information flow through the architecture proposed in the present chapter: following an encoding process, information passes through a convolutional network, leading to the eventual inference of the bacterial presence in a FOV.

6.3.2 Comparison of proposed and existing models

I compare my proposed method with a number of state of the art models from the VGG family [202], the ResNet family [98] (including Wide-Resnet [268]), the Densenet family [104], and InceptionV3 [222]. All models were pre-trained using the ImageNet dataset, which contains 1000 target classes and three input channels. Regardless of their initial configuration, each model's first convolutional layer was replaced with one that comprises a single input channel, kernel of size 3×3, stride value of 1, and padding of size 3×3. The alterations are motivated by the fact that the slide representation is monochrome (i.e. single channel) and the objects of interest are thin, elongated structures that frequently appear near the image boundary.

When transfer learning is used, the weights of the initial layer are summed to create a single matrix, which may have a negative effect on early training accuracy, but the benefit of not having to start training from scratch exceeds the impact on accuracy. The last modification is to the final linear layer, which is replaced with one that retains the same input features but has only one output node (in InceptionV3, this change is applied to its auxiliary classifier). The output weights of the last linear layer are passed through the sigmoid function.

6.3.2.1 Hyper-parameter learning

To ensure a fair comparison, all models, including the proposed one, are trained using the same set of hyper-parameters. To begin, the batch size is chosen to be 16 in order to achieve a balance between generalization, accuracy, and computing speed. Adam optimizer with the values of the β parameters (that is, the initial decay rates used when estimating the first and second moments of the gradient) equal to 0.50 and 0.99 is used. For the training process, following evidence from prior research [208], the base and maximum learning rates are set to 0.00001 and 0.0004, respectively, and the learning scheduler used is the novel circular scheduler with a step size equal to five times the size of the dataset (which varies according to batch size). Finally, binary cross entropy is used as the loss function; see Figure 6.3.

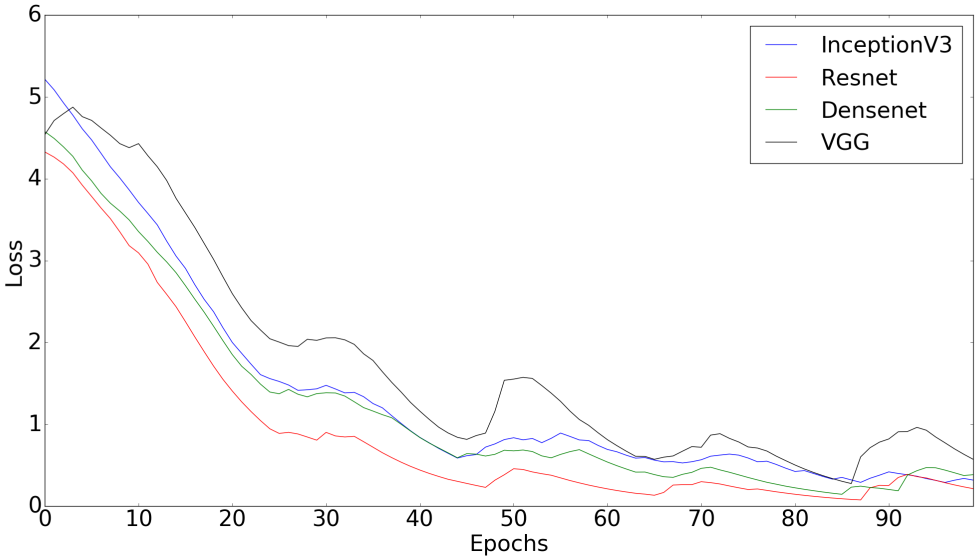


Figure 6.3: With the exception of VGG family, all models converge around the 65th epoch and exhibit the same overall learning behaviour. The poor performance of the VGG family suggests that the present task requires greater architectural sophistication than that achieved by merely stacking convolutional layers.

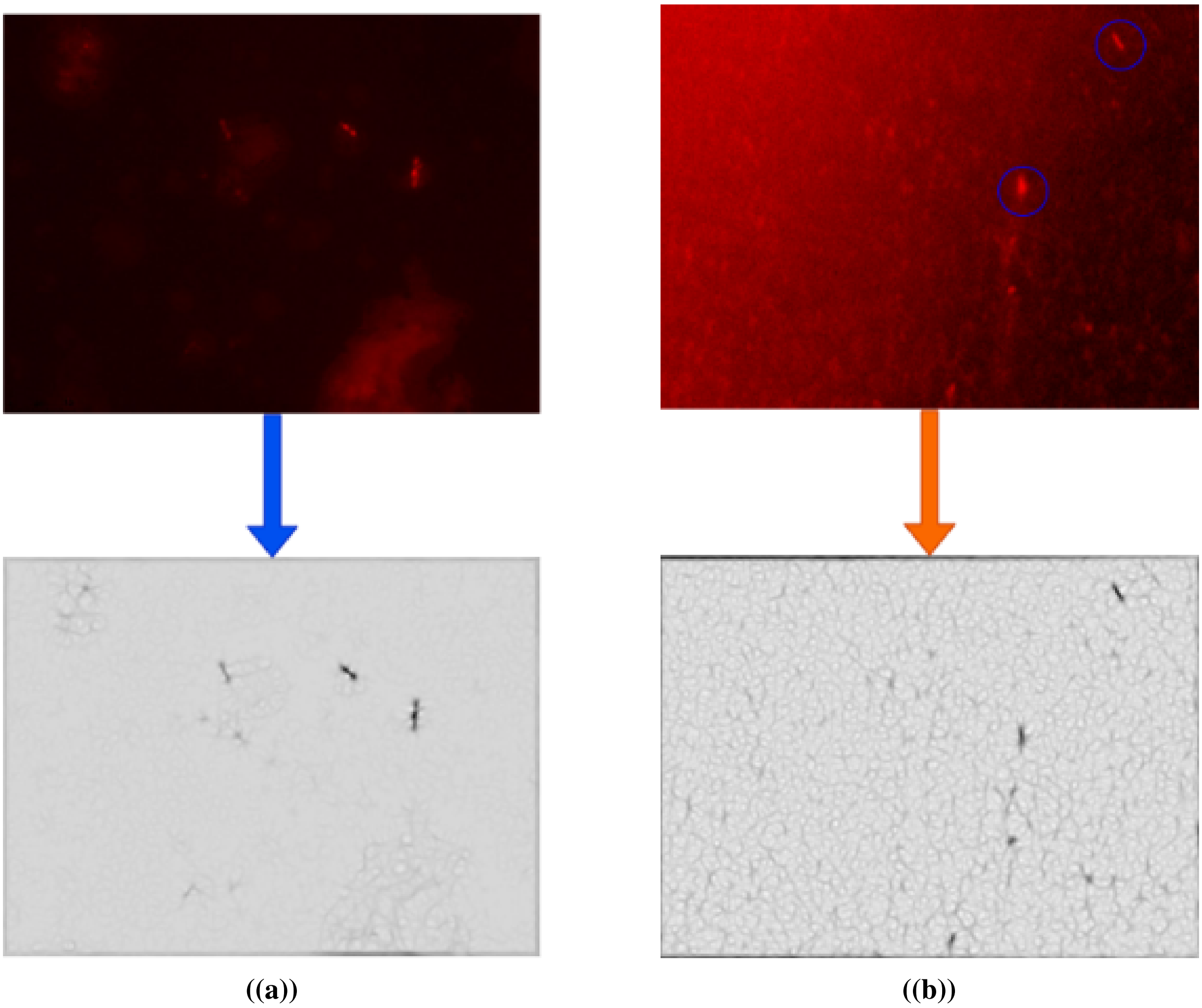
6.4 Evaluation

To assess performance of the FOV cropping and filter/classification method described in this chapter, the models introduced in Section 6.3 were evaluated by a series of metrics. As indicated in Chapter 3, consensus on standard classification metrics for this type of research is lacking; hence, I have chosen those which I deemed most suitable and explained my observations.

6.4.1 Data

The training dataset for the experimental work in this chapter, came from Image Set 1, which is described in Section 4.3.2. Around 800 positive FOVs were randomly chosen from Image Set 1 and 800 negative ones from Image Set 2. Moreover, 77 positive FOVs from Image Set 2 were incorporated into the training to enhance the models' adaptability to slides prepared differently. Figure 6.4 provides an illustration of FOV from the two sets. The reason behind this off choice is the fact that Image Set 1 does not contain any WSIs therefore no negative FOVs are present. To verify that the automated image analysis method being developed is not affected by changes in the morphology of *Mtb* cells during TB therapy, images were picked across all time periods of sample collection. These images were re-examined by a microscopist who had not seen them before and each FOV was categorised as positive (containing possible *Mtb* cells) or negative (not containing possible *Mtb* cells).

Evaluation of the method was done using 130 FOVs from Image Set 2, after performing the slide cropping procedures described in Section 6.2. A balanced sample of positive and negative FOVs was used. All FOVs were enhanced using the technique described in Chapters 4 and 5. Green and red channel FOVs were both used, even though microbiologists primarily use Auramine O staining in the green channel for bacterial detection: this is because red channel images, overall, are harder to evaluate with more difficulty to detect bacteria and worse background noise. Incorporation of the more difficult FOVs increase greater variability to the images examined by the network.



((a)) ((b))

Figure 6.4: Examples of various FOVs with LTR-red staining and their corresponding image enhancement representations. The first FOV **(a)** comes from Image Set 1 and the second **(b)** from Image Set 2. The difference in quality of the two FOVs is evident in their staining process, while the hardware and dyes employed are identical. On the second FOV **(b)**, the bacteria are scarcely discernible (marked with blue circles), but the image enhancement representation approach rectifies this issue.

6.4.2 Results

73 of the 130 FOV images in my test set were classified as positive (containing possible *Mtb* cells) and 57 were classified as negative by manual microscopy. Binary masks corresponding to all 130 FOVs, developed in collaboration with a microscopist, were treated as the ground truth for application of performance metrics. In order to facilitate a comprehensive comparison between models [237], I assessed performance using this ground truth labelling as the gold standard, using a number of metrics, namely overall accuracy, recall (sensitivity) and specificity accompanied by precision, ROC, and AUC.

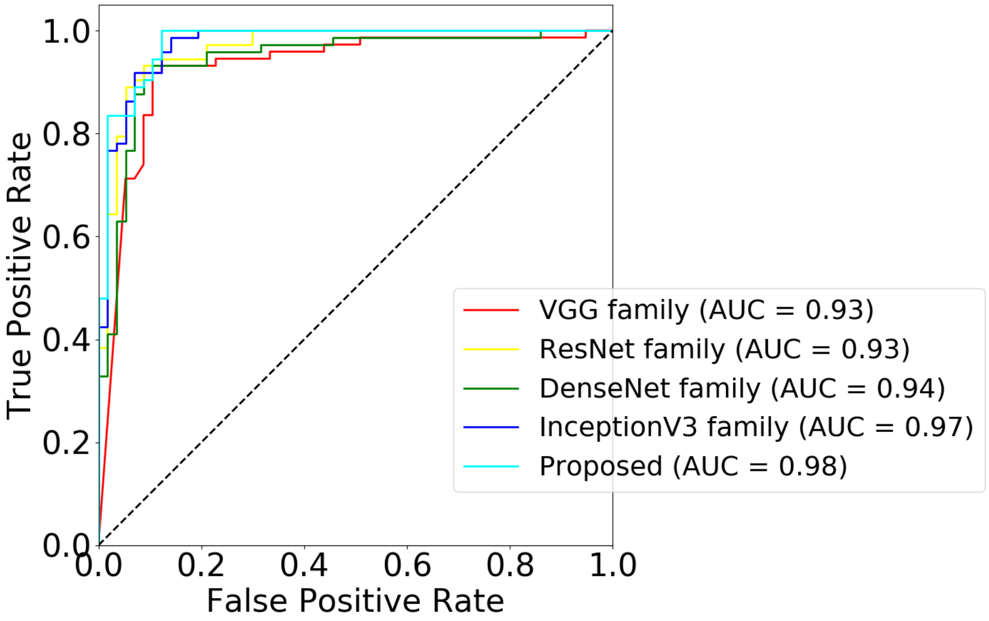


Figure 6.5: Comparison of ROC curves and the areas under the curves, both show that the proposed solution performs better on this test set from the current state of the art. For the sake of clarity, each model family is illustrated using the average performance of its evaluated models.

A summary of my experimental results is shown in Table 6.1. To start with, consider the overall performance metric in the form of the classification error (in the rightmost column of the table) and observe that the proposed method achieved the best performance of all 13 methods compared. The error rate of the next best model, namely ResNet50, is more than 11% greater. InceptionV3 and the best DenseNet family model, DenseNet201, performed next best (23% higher error rate than the proposed model). While there is significant variation between different specific models within all families, generally speaking ResNet performed better than DenseNet, and VGG networks fared the worst (45–101% higher error rate than the proposed model).

Table 6.1: Summary of results. All VGG models were trained using batch normalization (BN). The best performing model with respect to each statistic is shown in bold.

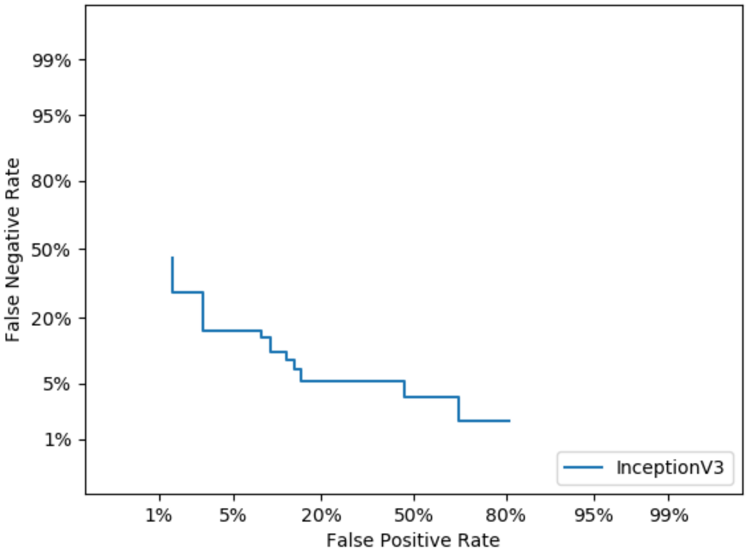
Model name	True +ve	False +ve	True -ve	False -ve	Sensitivity	Specificity	Error rate
VGG11 (w/ BN)	68	5	49	8	89.47%	90.74%	0.100
VGG13 (w/ BN)	67	6	47	10	87.01%	88.68%	0.123
VGG16 (w/ BN)	67	6	46	11	85.90%	88.46%	0.131
VGG19 (w/ BN)	63	10	50	7	90.00%	83.33%	0.131
ResNet18	65	8	51	6	91.55%	86.44%	0.108
ResNet34	66	7	51	6	91.67%	87.93%	0.100
ResNet50	67	6	**52**	**5**	93.06%	89.66%	0.085
ResNet50-Wide	68	5	49	8	89.47%	90.74%	0.100
DenseNet121	66	7	51	6	91.67%	87.93%	0.100
DenseNet169	67	6	45	12	84.81%	88.24%	0.139
DenseNet201	68	5	51	6	91.89%	91.07%	0.085
InceptionV3	68	5	**52**	**5**	**93.15%**	91.23%	0.077
Proposed	**73**	**0**	48	9	89.02%	**100.00%**	**0.069**

More detailed insight into the behaviour of different models can be gained by examining the specific error types (in columns 3 and 5 of Table 6.1; also see Figure 6.5). My proposed method performed best in terms of the false positive error rate – indeed, it made no incorrect positive calls at all. Other methods (e.g., DenseNet201, ResNet50, ResNet50-Wide, and VGG1) all had 7% false positive rates. On the other hand, my proposed method was not superior in terms of the false negative rate. In the context of this metric InceptionV3 and ResNet50 performed best, achieving the error rate of approximately 9%. Here it is important to consider the different effects of false positive (type I) and false negative (type II) errors on the task at hand [132]. A low type I error rate is useful as only salient FOVs (containing *Mtb* cells) are left in the dataset and large amounts of irrelevant information are removed, minimising bottlenecks for downstream analyses. On the other hand, type II errors, may result in FOVs containing some bacilli being missed. The incomplete detection of *Mtb* cells may be less consequential if sufficient information can be obtained from the identified cells. This may be less important if necessary information about *Mtb* cells can be gleaned from those which are identified, without every cell being detected but it may still be problematic if there are not many bacilli on the smear (so missing even a few could distort the data) or if the bacilli which are missed represent a cellular sub-population with a specific morphology which would have been important to see. Excessive type II errors may also adversely affect quantitative microscopy as they will result in systemic under-estimation of bacterial load. As illustrated by the results in Table 6.1. Type II errors do not represent a major challenge in this chapter, as most of the methods compared did not produce a high number of false negative results.

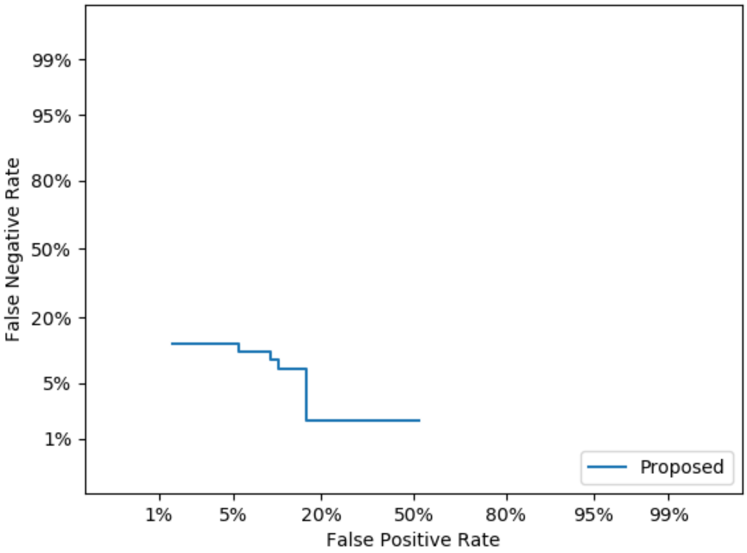
Overall, for different downstream clinical or research purposes differing levels of Type I and Type II can be acceptable. For each purpose, the choice about the model output might be modified to shift between positive and negative classifications. In other words, the model output without a sigmoid function constitutes a logistic regression, which in turn produces a probability. A logistic regression model that yields a value of close to 1 (e.g. 0.997) for a given FOV predicts that it is very likely to be positive. In contrast, an alternative FOV with a prediction score of close to 0 (e.g. 0.001) according to the same logistic regression model is very likely to be negative. Classifying an FOV becomes more ambiguous when its prediction score is 0.501, or any value proximate to the decision boundary. In order to map a logistic regression result to a binary category, it is necessary to determine a classification threshold (also known as the decision threshold). A value over the threshold is positive, whereas a number below the threshold is negative in this context. The output value of a sigmoid function falls within the interval $[0, 1]$, with the value 0.5 frequently serving as the classification boundary, however the specific threshold is subject to task-dependent adjustments.

Two methods can be adopted to manage results that are proximal to the decision boundary. Changing the class weights is a common technique used to develop models with a focus on minimising FN rate [204]. This was first developed as a solution for imbalanced data sets, to compensate for the under-representation of a certain class. Consequently, the most common threshold value (0.5) is maintained, but the logistic regression probability calculation differs owing to the weights applied to each class. For three reasons, I did not employ class weight balance throughout my training. Firstly, since the training dataset is balanced, adjusting class weights might not yield the intended benefit of facilitating decisions for ambiguous FOVs. Secondly, adjusting class weights introduces an extra hyperparameter to optimize, thereby increasing the model complexity. Class weights can sometimes make the optimization landscape more challenging, leading to slower or less stable convergence [88]. Finally, the difficulty with this strategy is that any improvement in accuracy may be accompanied by a significant decrease in sensitivity (recall) [204]. Some of the FOVs in both the training and test sets were subjectively assigned to each class by microscopists, who were uncertain of their objective classification. Consequently, favoring one class through weight adjustment could negatively affect model performance in this context. The efficacy of such weight modifications relies on accurate class specification, whereas in this case their allocation is sometimes subjective.

Changing the decision threshold is a common technique used to reduce instances of FN and FP. In order to decrease FN, the threshold value can be lowered since doing so will compel the model to predict fewer inputs as false, hence lowering the number of FN instances. Increasing the threshold value similarly reduces the amount of FP. I used a decision error trade-off (DET) graph to determine which threshold values should be evaluated [138]. A DET graph plots the FN rate vs the FP rate for binary classification algorithms [138]. Figure 6.6 shows a comparison of calculated DET graphs between my proposed model and the model with the second-best performance, InceptionV3. As the proposed model predicted no FPs, the approach taken was to compare models by minimising their FNs. In other words, the threshold is set lower than 0.5, which allows FPs to rise while decreasing FNs. I then compute the precision, recall, and F1-score for each threshold value to determine the effect of the trade-off between the two models; see Table 6.2. The proposed model exhibits a less significant decline in performance upon adjustment of the decision boundary than the model with the second-highest performance does.



((a))



((b))

Figure 6.6: DET graphs of InceptionV3 **(a)** and proposed model **(b)**. Graph **(a)** covers a larger region than graph **(b)**, indicating that the predicted effect of the trade-off will be greater.

Table 6.2: Summary of results with different threshold values. The F-1 score comparing the two models indicates that altering the threshold boundary for InceptionV3 incurs a bigger penalty than for the proposed model. In other words, for every reduction in FN, InceptionV3 pays a larger rise in FP numbers than the proposed model. Therefore, the proposed model is well tuned and offers flexibility in error type balancing (type I and type II)

Threshold value	Precision	Recall	F1-score	Precision	Recall	F1-score
	InceptionV3			Proposed		
0.5	0.94	0.92	0.93	1	0.89	0.94
0.2	0.95	0.87	0.91	1	0.91	0.95
0.01	0.97	0.84	0.90	0.94	0.94	0.94

6.5 Discussion

The work described in this chapter shows that use of WSI of TB microscopy slides, stained using the dual Auramine O LTR technique, is possible. The WSI which is generated can be cropped into FOVs, which in turn are suitable for assessment using DL tools to filter those which contain possible *Mtb* cells from the much larger number of fields which contain only background.

My proposed method for FOV classification learns from coarsely labelled images and their corresponding binary masks and outperforms other generic CNNs based on standard performance metrics. Furthermore, the proposed model offers an additional advantage, as demonstrated by decision threshold trade-off analysis. Different applications of automated smear microscopy may require different performance characteristics, e.g. a clinician aiming to simply classify whether an entire smear microscopy slide is “positive” or “negative” is likely to be more tolerant of Type II error (false negative cell detection) than a researcher wishing to accurately count the bacterial load or describe subtle changes in diverse bacterial populations within a sample. With my model, the penalty for decreasing Type II error, at the expense of higher Type I error was smaller than the second best competitor model. This gives clinical and academic microbiologists greater latitude to adapt my method for satisfactory completion of their intended task.

The potential value of this approach is that automated pre-selection of *Mtb*-containing FOVs could accelerate creation of suitable image sets for advancement into the DL cell counting or phenotyping models described in Chapters 5 and 7. However, there are some limitations. The slide scanner and associated software and computing power required to generate WSIs is expensive, probably limiting its use to well-resourced research laboratories in the first instance. The FOV classification method introduced in this chapter is also dependent on shape characterisation, a fundamental aspect of which involves generating feature maps from manually created binary mask counterparts of the input images, both for training and testing. This need for human labour to do this still conflicts with the primary objective of this thesis chapter, which is to automate as much of the smear microscopy process as possible as much as possible. In Chapter 7 I will partly address this limitation by revisiting a different approach to bacterial detection through semantic segmentation. The ultimate goal is not only to detect bacteria but also to automate the creation of essential binary masks, as required by methods like the one proposed in this chapter.

So far, in this thesis I have demonstrated the ability of AI tools to enhance sputum smear microscopy images, to detect and quantify *Mtb* bacilli in pre-selected FOVs and to filter those FOVs out of large dimension WSIs. However, it will be recalled from Section 2.7 that an important TB therapeutic research goal is still to investigate and monitor treatment response by studying changes in the size and lipid content of *Mtb* bacteria in different growth conditions and under drug pressure [7, 207]. The next, and final, chapter of my thesis will focus on the development of TB-AI microscopy methods to achieve this.

Chapter Seven

Estimating Phenotypic Characteristics of Tuberculosis Bacteria

Chapter Abstract– In Chapter 1, the possible importance of intracellular lipid content and cell dimensions as phenotypic characteristics of *Mtb* bacilli was explained. Here, I seek to automate the process of examining sputum smear microscopy FOVs in order to determine those characteristics at the individual cell level and facilitate further research into their importance. I propose a UNet-based model to rapidly localise potential bacteria inside a FOV. I introduce a novel method that uses Fourier descriptors to exclude contours that do not belong to the class of bacteria, hence minimising detection of false positive objects. Finally, I propose my own feature extractor in conjunction with feature descriptors as a means of extracting a representation into a support vector multi-regressor in order to estimate the length and width of each bacterium. Using Image Set 1 the proposed method i) outperformed previous methods for the bacterial detection task by almost 8% (SD coefficient) and ii) estimated the cell length and width with a root mean square error of less than 0.01%.

7.1 Chapter introduction

As outlined in Chapter 2, one of the most compelling arguments for the continued use of smear microscopy in the context of TB treatment monitoring and therapeutics research is its capacity to investigate changes occurring within individual *Mtb* cells throughout the course of therapy. Recall from Section 2.7.2 recent microbiological research suggest that the physical morphology of each organism offers phenotypic information on its physiological behaviour in relation to antibiotic susceptibility. For example, some *Mtb* cells accumulate nonpolar lipids intracellularly, allowing them to be classed as LR rather than LP [97, 53, 80, 96, 57]. In vitro data suggest that LR bacteria are antibiotic tolerant (less easy to kill by the first-line drugs used to treat TB) [96, 57] and may play a role in poor patient outcomes (treatment failure or post-treatment relapse) [207]. The Auramine O LTR staining method, which is used in both Image Sets analyzed throughout this thesis, allows discrimination between LR and LP cells [80, 207].

Additionally, in vitro microscopy has previously demonstrated that *Mtb* cells grow asymmetrically, creating variation in cell length over time [39, 7]. Cells of different sizes with different growth poles have variable susceptibility to individual antibiotics [7, 184]. Preliminary clinical data suggest that the median length of persistent *Mtb* cells may be associated with worse disease severity and increases after antibiotic exposure [243, 19].

To understand whether phenotypic changes in *Mtb* bacilli, e.g. variable intracellular lipid content or cell dimensions, really are useful characteristics for the study of TB treatment response, larger scale laboratory and clinical studies are required. However, (as repeatedly observed in this thesis) smear microscopy is time-intensive and subjective which makes this work difficult to perform at scale [185]. Each slide must be examined in discrete FOV that are inspected sequentially. This process is tiring which can introduce errors [185]. Some slides are challenging to evaluate because AFB might have odd appearances or because non-bacterial components (artefacts) inside the sputum matrix mimic *Mtb* cells. These problems have been noted in previous chapters in relation to classifying or counting bacteria but are even more challenging when additional properties of individual cells must be evaluated one by one. Application of contemporary DL techniques may help tackle the issue of cell phenotyping [241]. Recent studies have demonstrated significant accomplishments in the realm of automated diagnosis, treatment monitoring, and the potential prevention of other medical conditions (e.g., cardiovascular and gynaecological pathology) [256, 220].

In this chapter I aim to advance ML-based approaches to phenotyping *Mtb* cells from fluorescence microscopy images by developing methods to:

- Locate *Mtb* cells within given FOVs on Auramine O and LTR stained fluorescence microscopy images, with performance evaluation by two established metrics (Jaccard index and SD coefficient).

- Co-localise the same *Mtb* cells on paired Auramine O and LTR stained images of an FOVs, in order to assess the proportion of LR bacteria.

- Estimate the length and width of *Mtb* cells in FOV patches from sputum smears collected at 0, 2 and 6 months of therapy.

7.2 Related work

Whilst prior research on automation of smear microscopy for TB diagnosis was extensively reviewed in Chapter 3, I was unable to find any other work that sought to studying morphological phenotypes of *Mtb* cells in the context of treatment response. However, some previous literature describes use of DL tools for phenotypic evaluation of other cell types in order to understand the pathophysiology of infectious diseases and bacterial response to antibiotics.

In the realm of mycobacteria, Bao *et al.* used using light microscopy and convolutional neural networks (CNN) to classify morphological alterations of macrophages infected with Mycobacterium marinum, a surrogate model for *Mtb* and a pathogenic bacterial species in its own right, to show the role of the essential virulence factor EsxA [17]. Whilst this work focussed on identification phenotypic changes in host cells rather than bacteria and did not fall under the purview of treatment monitoring, it still demonstrates the capacity of automated image analysis to detect changes in cell appearance of individual cells which enhance our understanding of bacterial pathophysiology.

In the domain of antibiotic response, Yu *et al.* assessed susceptibility of Escherichia coli bacteria in urine to five relevant antibiotics using DL video microscopy [258]. Conventional procedures for antimicrobial susceptibility testing can take several days and delay clinical decision making, but these authors described a technique that used a 7 layer CNN to evaluate footage of freely moving bacterial cells in real time. Inhibition (or not) by antibiotics was reported by learning several phenotypic characteristics of the cell without requiring the definition and quantification of each characteristic. Antibiotic susceptibility was reported with mean accuracy of 91.8% within 30 minutes. Similarly, Zahir *et al.* used high throughput screening and DL to describe phenotypic 'bulging' in E. coli which is associated with resistance and tolerance to β-lactam antibiotics [269].

7.3 Proposed method

The methods proposed in this paper consists of three stages: i) bacterial detection from microscopy FOVs, ii) paired detection of bacterial locations from two images of each FOV (one captured to show Auramine O staining of *Mtb* cells, and one captured to show LTR staining of intracellular lipid; collectively these allow inference of the proportion of LR bacteria in the FOV), and finally iii) estimation of individual bacterial dimensions (length and width) from cropped patches of FOVs containing one or more *Mtb* cell. Segmentation techniques are used for stage i) and ii), and regression is used for stage iii). These methods are designed and evaluated separately, with distinct objectives and evaluation criteria. Although I describe them to operate independently, they could be used sequentially with a future goal of pipeline integration.

The initial stage involves revisiting the method of semantic segmentation, employing a modified version of the widely recognized UNet [188] architecture, while the regression component utilizes a specialized CNN.

7.3.1 UNet: segmentation-based CNN

As explained previously, CNNs leverage convolutional layers to extract hierarchical features from images. A UNet model builds on these CNN principles to create an encoder-decoder segmentation architecture [188].

The encoder portion of UNet uses repeated blocks of convolution, activation, and max pooling layers, similarly to a typical CNN. This encodes the input image into high-level feature representations while downsampling spatially. The decoder pathway then upsamples these features back to the original input resolution using transpose convolutions. A key difference from a CNN is the introduction of skip connections that concatenate encoder features with the upsampled decoder features. These skips provide the decoder with both contextual information from the encoder (information recall) as well as fine-grained localization from the upsampled features.

Finally, the decoded features are fed into a convolution layer to generate a pixel-wise probability map for semantic segmentation. So UNet leverages a CNN encoder to analyze contextual features, but adds a decoding path with skip connections to localize and precisely segment input images in an end-to-end manner. The model is trained via backpropagation just like ordinary CNNs. This architecture remains popular for segmentation tasks, especially in medical imaging where precision is critical.

In summary, UNet extends CNNs into an efficient encoder-decoder structure specialized for precise pixel-level segmentation while retaining automated feature extraction capabilities. Figure 7.1 provides a visual example of the UNet architecture.

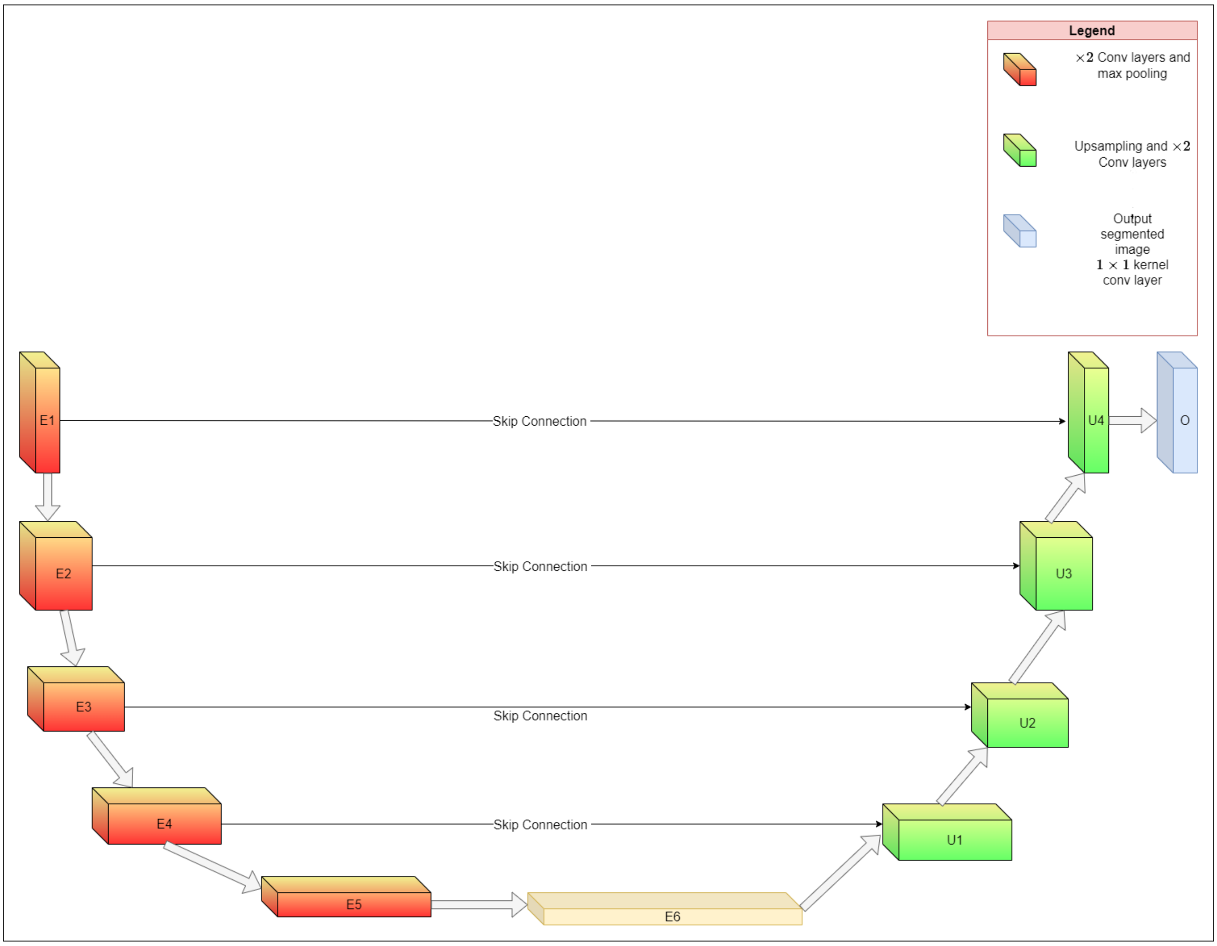
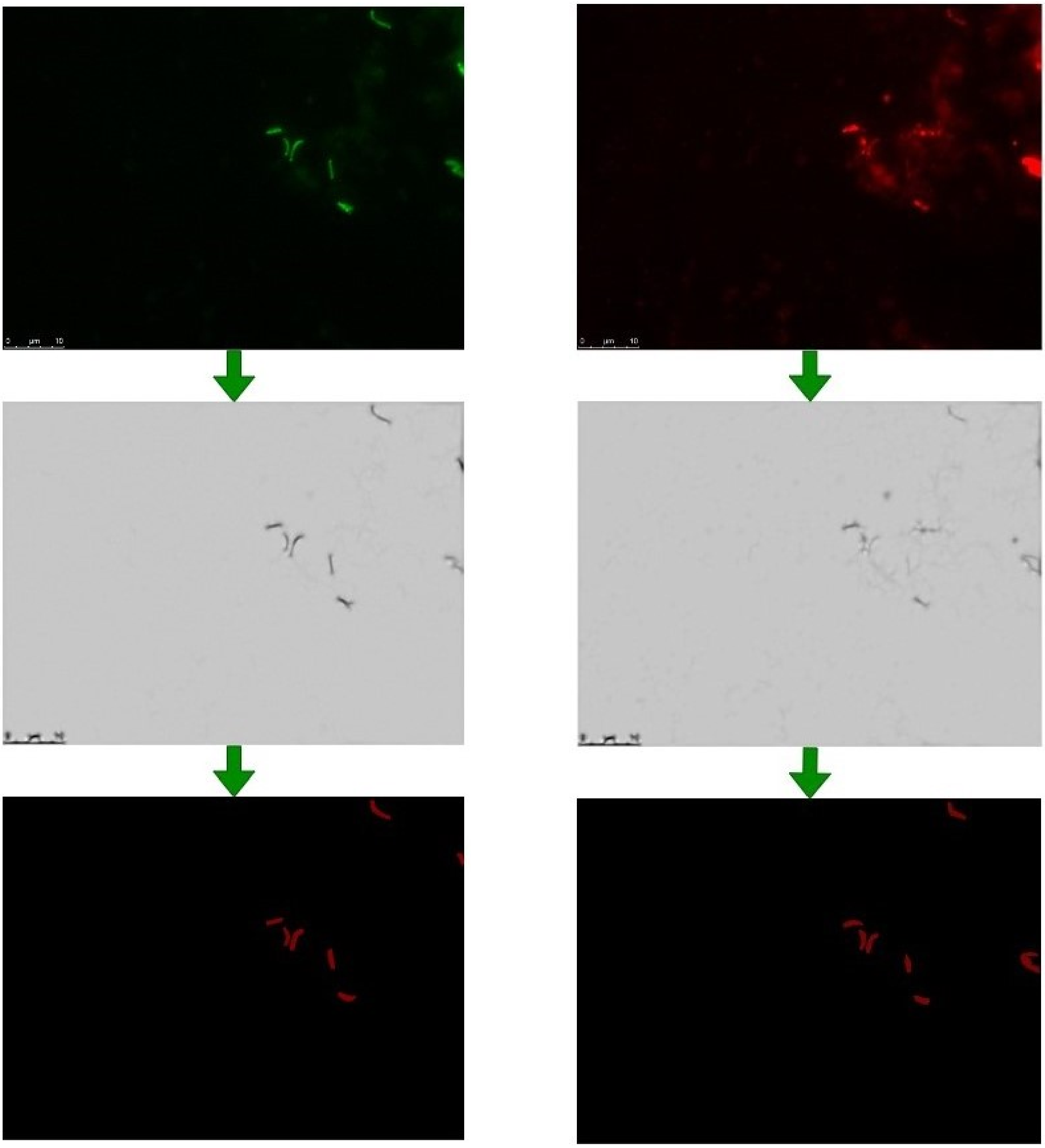

Legend
×2 Conv layers and max pooling
Upsampling and ×2 Conv layers
Output segmented image 1 × 1 kernel conv layer
E1
E2
E3
E4
E5
E6
U1
U2
U3
U4
O
Skip Connection
Skip Connection
Skip Connection
Skip Connection

Figure 7.1: Flow of information of the UNet architecture.

7.3.2 Bacteria detection

As described in Section 2.7.3.2, lipid content within *Mtb* cells is calculated as the proportion of total bacteria detected in an FOV stained with Auramine O (green channel) which are also detected in the same FOV stained with LTR (red channel) at the same location. As each FOV is represented as two RGB images, I set the red and blue channels to 0 to make bacteria visible in the green channel only. Similarly, I set the green and blue channels to 0 makes bacteria visible in the red channel only. If a bacterium is localised in the green channel and co-localised at the same spot in the red channel, it is LR. If the bacterium is localised in the green channel without co-localisation in the red channel, it is LP. If no bacterium is localised in the green channel, any object identified at that spot in the red channel is discarded as artefact. This is because Auramine O is a gold standard microscopy stain for *Mtb*, whilst LTR stains lipid irrespective of whether it is within bacteria of interest. For example, when examining paired images of an FOV, if 5 bacterial locations are found in the green channel, 3 are co-localised in the red channel only, 5 *Mtb* cells have been identified in total, 3 (60% lipid content) of which are LR and the other 2 are LP. Any other objects resembling bacteria in the red channel are also discarded, as microbiologists typically address this task in a unidirectional manner rather than bidirectionally, i.e. first detect bacteria in the green channel and then the red one.

The key component in this analysis is not the actual colour intensity, but rather the intensity of the object in contrast to the image background when the other two channels of an RGB image are set to a value of 0, i.e. suppressed. I convert each FOV to greyscale in order to avoid the need for extra training for each coloured FOV and also to reduce complexity when moving from three dimensions to one. In addition, to make FOVs less susceptible to noise, I use the image enhancement technique described in Chapters 4- 6. Figure 7.2 presents an example of the pre-processing procedure, as well as featuring segmentation ground truths.



((a)) ((b))

Figure 7.2: Examples of paired images from the same FOV from Image Set 1. In the top row (a) all colour channels except green suppressed and (b) all colour channels except red suppressed. In the second row, all images are converted to grayscale using the proposed image enhancement technique. In the bottom row, manual ground-truth labelling of cells in the two images is shown; the two separate fluorescence labels visible on the same FOV display different information, necessitating a different ground truth label. At this phase, I train the model to detect as many *Mtb*-shaped objects as possible in both green and red channel images, even although objects which are only detected in the red image are ultimately discarded.

The first objective is to binarize the FOV, which means that the resulting image has a black background and the objects of interest (bacteria) are white areas. I adapt and apply UNet [188] since it is an effective choice for learning to collect important information about objects of interest and generate a binarised image. I replace the first layer of the UNet with one that has input channels of one rather than three and output channels of 32 rather than 64 as required by the original implementation. Therefore, the input and output channels of subsequent layers are adjusted to align with the original UNet implementation, whereby the number of channels in each layer is doubled compared to the previous layer. Consequently, the proposed network architecture exhibits a reduction in the number of channels at the bottleneck level from 1024 to 512. Kernel sizes and padding for the convolutional layers are not changed. In addition, the max pooling layers in the model have a stride of 1, as opposed to the original UNet that used stride of 2, while the kernel size remains the same. These modifications of the layers are driven by the fact that bacteria do not have complicated shapes. As the form of a bacterium is relatively simple, the first layer requires less deductive reasoning; therefore, higher channel layers may cause the model to overfit on the training data and acquire extraneous features. As this is supervised learning, each FOV used to train the network has previously been examined by an experienced microscopist who has manually highlighted bacterial outlines in each FOV, which is then converted into a binary image and used as ground truth for both the UNet and the proposed network training.

7.3.2.1 Training of segmentation networks

The training is carried out in an end-to-end fashion; there is no use of transfer learning. Due to the fact that there is no transfer learning, I train the network for over 1000 epochs. Similar to the approach taken in Chapter 5 during the training of the Cycle-GAN, I opt to use AdaBelief [274] for training the UNet in this context as well.

A circular scheduler with a step size equal to five times the size of the dataset (which in turn is dependent on the batch size) is used in conjunction with a learning rate of 0.0001, which is the default setting [208]. Both the base learning rate and the upper learning rate are set to their respective default values of 0.00001 and 0.0004. I use Dice loss [126] (also known as F1-score) as the loss function to train the model. To increase the robustness and generalizability of the learning process, I augment real data with synthetic data. To achieve this I synthesise images randomly rotated by ±25° and mirrored around the vertical or horizontal axis; this increases the quantity of training data by roughly 50% [260]. Note that this type of enhancement is particularly well suited to the task at hand because, unlike natural images, in which there is an inherent asymmetry in directions (e.g., the horizontal and vertical directions are objectively defined and cannot be swapped), in the microscopy slides of interest, all directions are interchangeable and therefore equivalent. Furthermore, input images are resized to 256×256 pixels using bicubic interpolation [8].

7.3.2.2 Minimizing false positives with bacterial morphological features

Existing literature, and some results from Chapters 4 to 6, describe that detection of bacteria using gradient-based methods alone is not always successful [174, 137]. Specificity can be compromised by false positive misinterpretation of artefacts as bacteria, on the basis of similar colour intensity. To reduce detection of these false positives, my method includes heuristic morphological characteristics (area, perimeter, number of edges, and Fourier descriptors). Utilizing the Douglas-Peucker technique [67], I determine the contour's area, perimeter, and approximate shape, employing the same approach as detailed in Chapter 4. Essential parameters for a detected shape to be identified as a bacterium are: the area must be between 80 and 1200 pixels, the perimeter must be between 40 and 300 pixels, and the approximate form must have between 9 and 20 edges.

In the last step of this process, I calculate the elliptic Fourier descriptors for each contour of the ground truth labels using the 20th harmonic. The application of the 20th harmonic for representation yields proximate coefficients that capture well the morphology of a majority of the designated bacteria specimens chosen at random; see Figure 7.3. Higher numbers of harmonic result in an overfitted outline of the current contour. Once each Fourier descriptor for every contour has been computed, the resulting matrix has the dimensions $n \times 20 \times 4$, where n is the total number of contours. The last dimension, 4, reflects the coefficients returned, of the Fourier series representation of the contour. The final 20×4 matrix is created by averaging the Fourier descriptors from all calculated contours. Furthermore, the Fourier descriptors of each predicted contour are calculated. These descriptors are then used in the calculation of the Euclidean distance between the average descriptors derived from the ground truth labels. To be considered a valid bacterial shape, the Euclidean difference between the predicted contour Fourier descriptors and the average descriptors must be between 14 and 18 pixels.

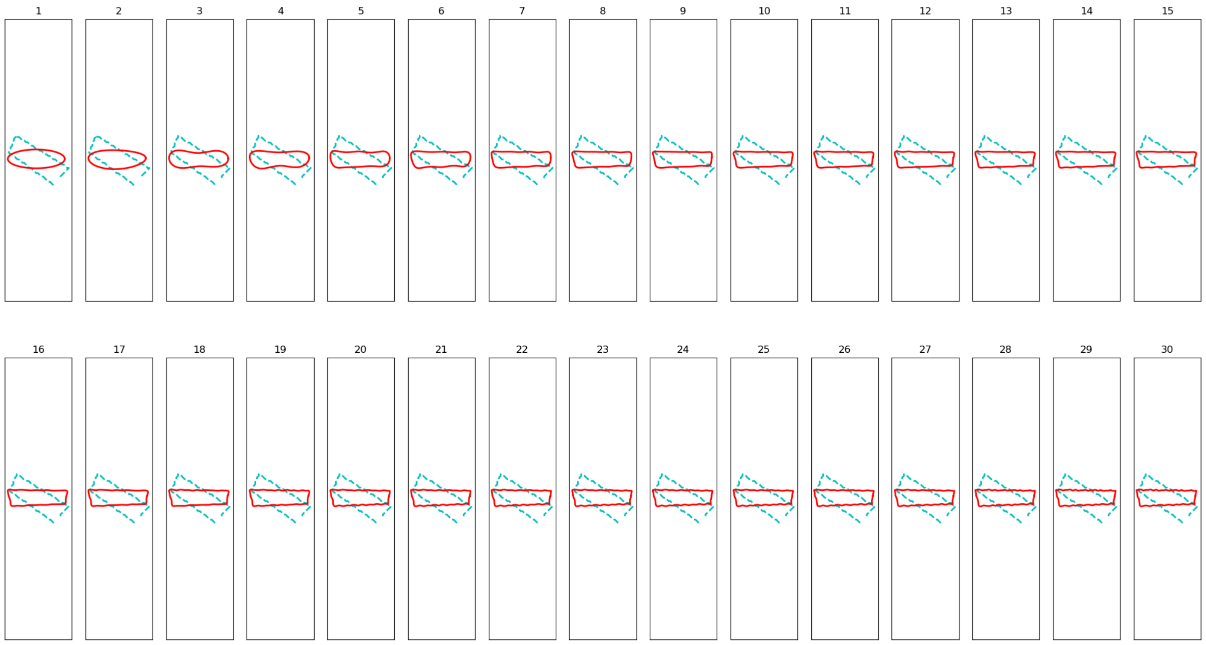


Figure 7.3: Several plots depicting the impact of the harmonic value. Lower numerical values generate a shape that is relatively generic, whereas higher numerical values endeavour to achieve an exceedingly precise match with the shape.

7.3.3 Estimating cell length and width

For the last step of this process, any bacterium/contour in the green channel images that fits the requirements given in Section 7.3.2.2 is utilised as the test set. Firstly, a microscopist manually crops patches containing one or more bacteria that overlap and annotates the cells with straight lines down their entire length. Multiple straight lines are needed for bacteria with curved or angular forms. Since the cell width across all cells is very similar (typically 5-6 pixels), width is averaged per patch, thus a patch with three cells is represented by a scalar for its width. Furthermore, I observe that the maximum number of bacteria per patch is four (n.b. for my dataset), the size of the vector acting as the ground truth label during training is five. If a patch contains two bacteria, for example, the first entry is the average width, the second entry is the sum of the lengths of the first bacterium, and the third entry is the sum of the lengths of the second bacterium. The remaining entries are all 0. Evidently, an additional benefit of this approach is its ability to count the number of bacteria present in a patch, similarly to previous work [241] and Chapter 5.

I utilise these labels to train a second CNN model, using regression, i.e. the final linear output layer does not contain a sigmoid activation. The trained model is stored and later deployed as a pre-trained model with its linear output layer removed, transforming it into a feature extraction encoder of 128 sized vector. Additionally, several feature descriptors are applied to extract a supplementary 128 sized vector of features from the input patches. These are: RootSIFT [9], Multiple Kernel local descriptors [150], HardNet [144], HardNet8 [180], HyNet [227], TFeat [15], SOSNet [227], Histogram of Oriented Gradients[1] [52]and Local Binary Patterns [158, 72]. The two vectors, one from the CNN and the other from the feature descriptor are then concatenated, creating a 256-dimensional feature vector. This vector serves as input to a multi-output support vector regressor (MSVR) [16] aiming to predict the same 5-dimensional ground truth. Figure 7.4 shows an overview of the method's information flow.

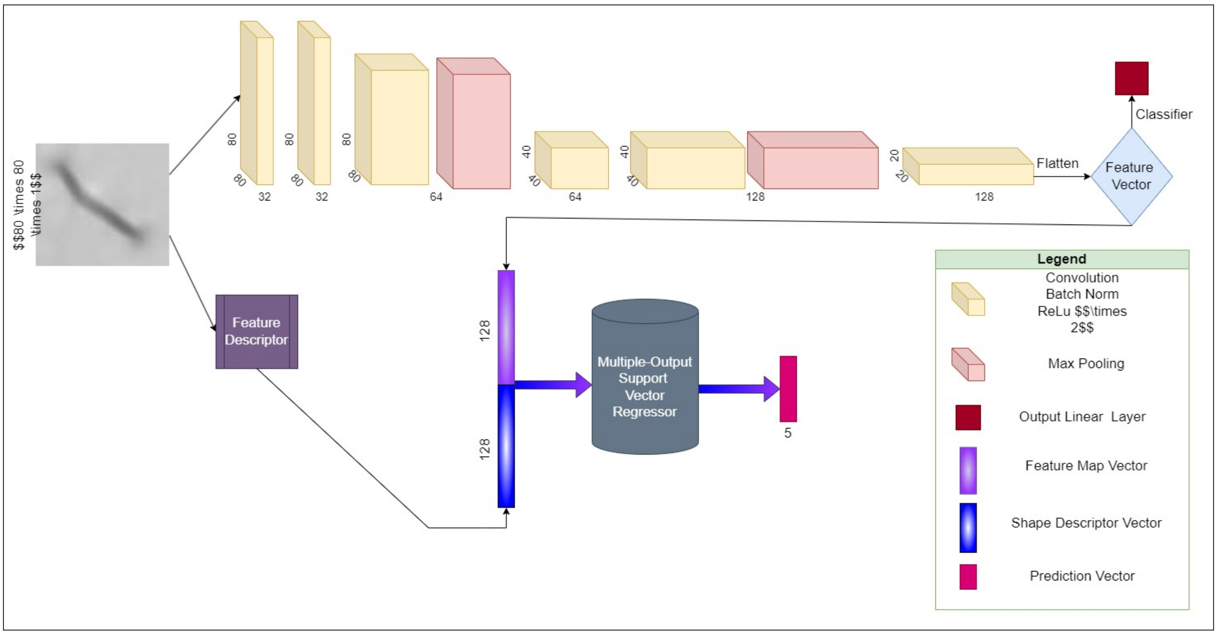


Figure 7.4: Following an encoding procedure from both the CNN and the feature descriptor, the MSVR outputs the final predictions. Since 1×1 convolutional filters may be used to modify the dimensionality of the filter space while maintaining linear activation of pixel values, the kernel size of all CNN layers is set to 1. This is due to the small dimensions of the input images; thus, I want to capture bacterial characteristics without losing spatial information.

7.3.3.1 Training setup

As with the training for segmentation, no transfer learning is performed in this instance, and the model is trained from scratch for 1,000 epochs. For the optimiser I use Adam for its straightforward implementation, being computationally efficient, and low memory requirements. The hyper-parameters β_1 and β_2 are set to 0.5 and 0.999 respectively, the learning rate to 0.001, and the cosine annealing learning rate scheduler employed [135]. This scheduler decreases the learning rate every 20 iterations until it reaches 0.0001 before initiating over again. Finally, the

[1] Given that input patch size is 80×80 pixels, to output a 128-d vector, orientation bins is set to 8, pixels per cell is set to 20×20 and cells per block is set 1.

loss function used is the Least Absolute Deviation (L1), since the dataset contains many outliers which are emphasised by squared differences. Considering that I have 1000 patches available for training in this stage (80% for training and 20% for testing), no data augmentation is performed. Like the previous CNN, the input patches are made square before being scaled to 80×80 pixels using bicubic interpolation. Following grid search hyperparameter learning, the following are used: a radial basis function (RBF) kernel, $C = 1$, $\epsilon = 0.001$ and $\gamma = 0.01$. Figure 7.5 presents a graphical summary.

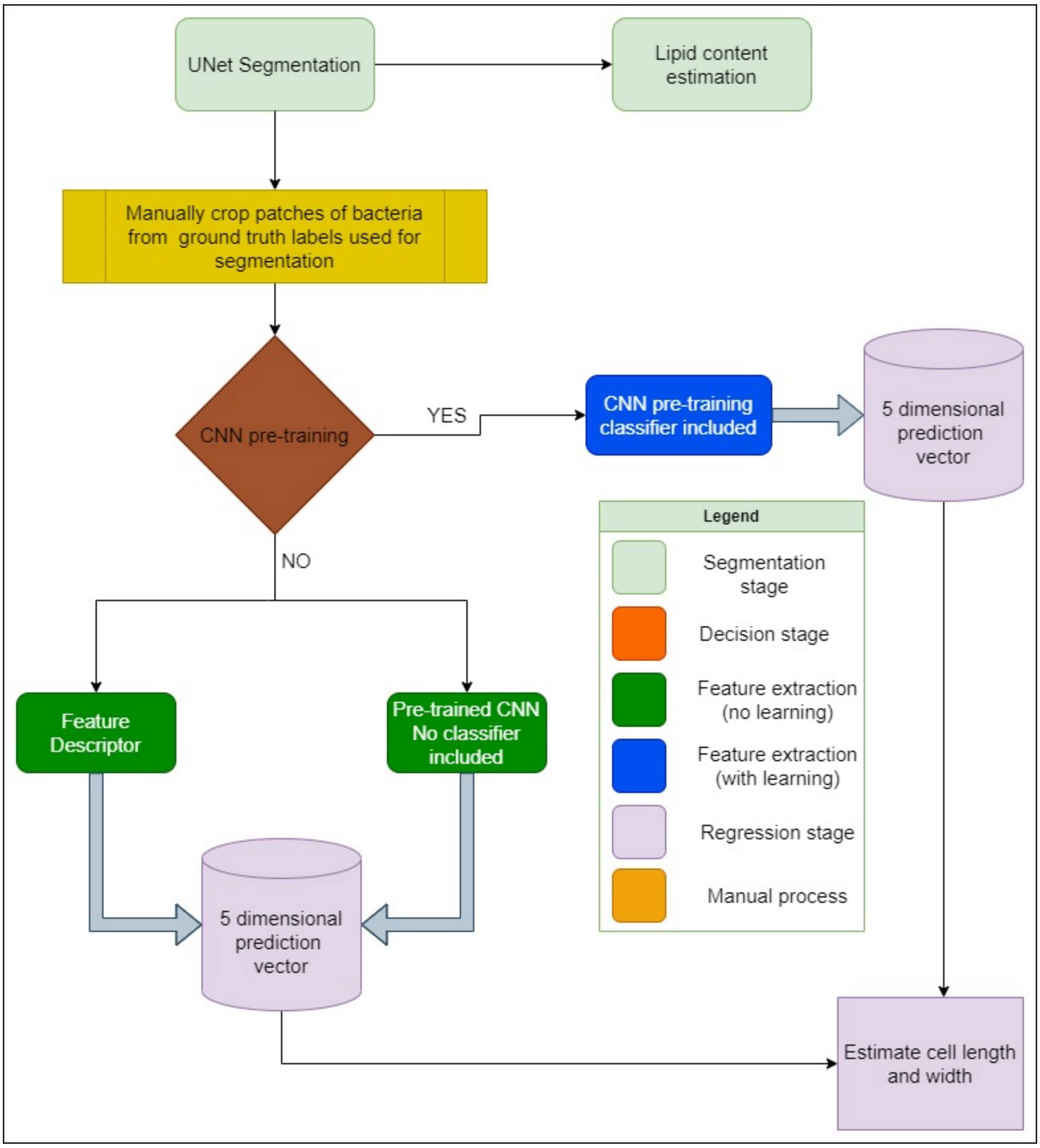

UNet Segmentation
Lipid content estimation
Manually crop patches of bacteria from ground truth labels used for segmentation
CNN pre-training
YES
CNN pre-training classifier included
5 dimensional prediction vector
NO
Legend
Segmentation stage
Decision stage
Feature extraction (no learning)
Feature extraction (with learning)
Regression stage
Manual process
Feature Descriptor
Pre-trained CNN No classifier included
5 dimensional prediction vector
Estimate cell length and width

Figure 7.5: The diagram presented illustrates the proposed approach. The first method involves the training of the UNet model and the proposed network. Afterwards, bacterial patches are manually cropped from ground truth labels in the green channel that were employed in the segmentation method. The decision stage involves determining whether it is necessary to pre-train the CNN model for the final stage, or alternatively, to utilise the same CNN along with a regression layer to make predictions on the vector representing cell length and width. If pre-training is not needed, the features extracted from the pre-trained CNN are combined with the output of a feature descriptor. This concatenated feature representation is then fed into a MSVR, which produces a prediction vector that resembles the output of the CNN's regression layer.

7.4 Results

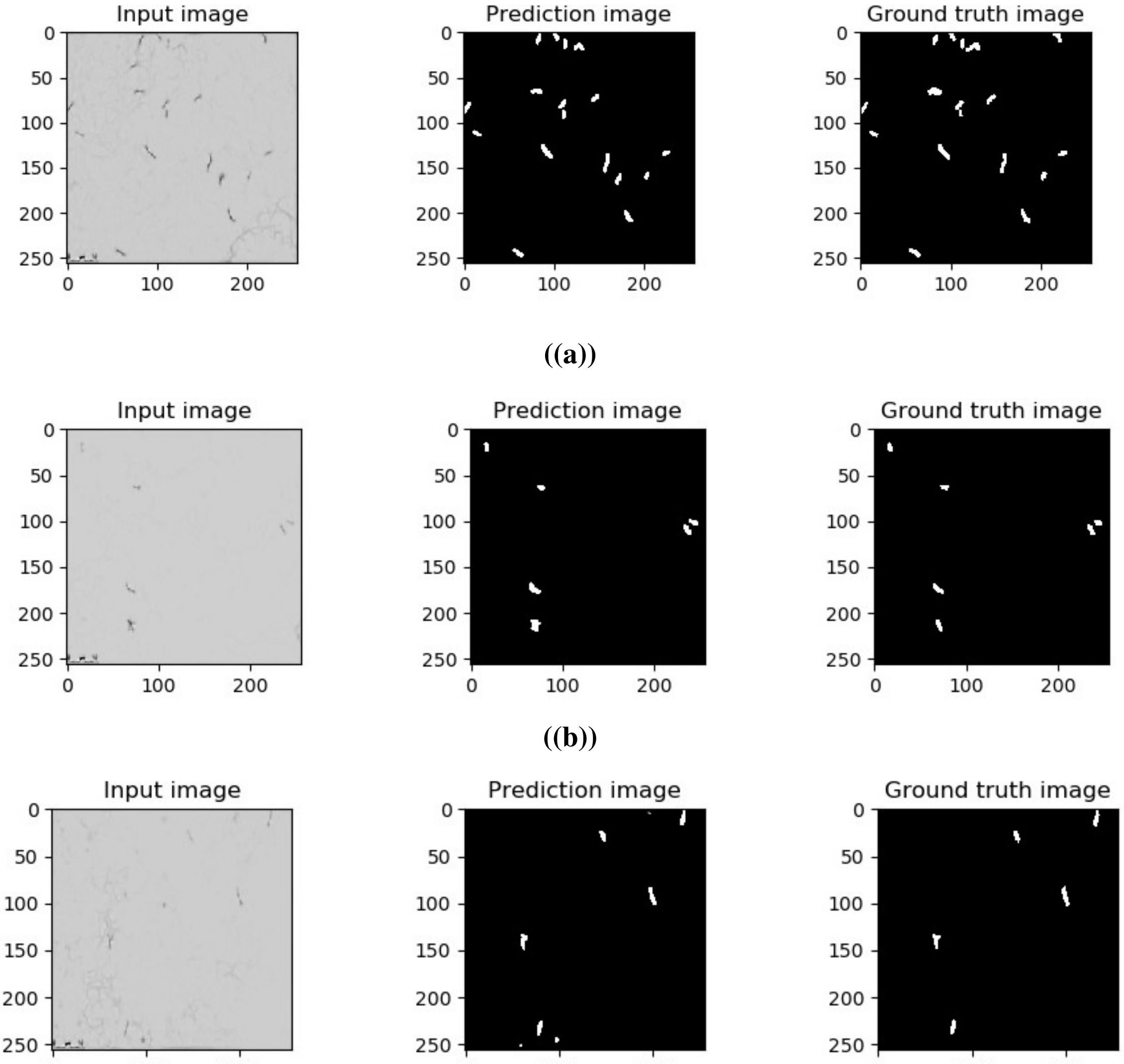
In this section, I will describe an empirical assessment of the proposed algorithm using data from Image Set 1. I will start by providing a description of the data used and then proceed to evaluate each method separately, taking into consideration the specific context of each.

7.4.1 Dataset

In all experiments conducted in this chapter, Auramine O and LTR FOV from Image Test Set 1, as detailed in Section 4.3.2, were employed. A selection of 500 FOVs pairs, spanning all time points of sample collection in the original clinical study, was made to ensure that the automated detection network for *Mtb* bacteria, remained unaffected by potential alterations in bacillary morphology during the course of TB treatment. To create ground truth images for the segmentation analysis, a microscopist who was independent of the original project which generated these images re-examined them, labelling objects of interest in both the green and red channel images. Importantly, in contrast to the labeling approach employed in Chapter 5, the annotation procedure involved outlining different bacteria rather than using bounding boxes; for visual reference, see Figure 7.2.

7.4.2 Semantic segmentation of bacteria detection

As explained in Section 7.3.2, bacterial detection and estimation of lipid content must be done in combination. Therefore, evaluating the performance of these tasks should be done together. However, distinct techniques are required to assess the separate processes of semantic segmentation on green and red channel images of an FOV, and distance-based evaluation of whether the same objects have been localised on both images. For example, although being a true positive in terms of detection, for the lipid content it cannot be deemed accurate. This is also the primary reason why these two stages of this work require two distinct assessment techniques.



((a))

((b))

((c))

Figure 7.6: (a) and (b) are separate examples of Auramine O stained FOVs, the prediction image from the segmentation model and labels applied by a microscopist to corresponding ground truth images. (c) is an example of a different LTR stained FOV (not paired with (a) or (b)). The prediction in (c) has localised three false positives objects, which are likely due to noise or artefacts and are subsequently rejected using the morphology-based approach.

The performance metrics used in the assessment of semantic segmentation are the SD [93] and Jaccard index [21]. When only Auramine O (green channel) stained FOVs are included, these are 97.00% and 96.06% respectively for the test set. The value of the Jaccard and SD coefficient index exceeds that achieved by earlier efforts [60, 147, 209]. All works employ the same evaluation metrics, which facilitates direct comparison with this method. However, when LTR (red channel) stained images are included, the percentages decrease to 92.03% and 85.84%, respectively. As seen in Figure 7.6, LTR stained FOVs often result in false positives, which motivated the subsequent use of morphology characterisation. Appendix C.1 contains further illustrative comparisons between the segmented Auramine O and LTR FOVs. A comprehensive comparison of the outcomes from the two models (UNet and proposed), appears in Table 7.1. Following the application of morphological criteria and Fourier descriptors to the FOVs of both dyes, the final percentages are 95.47% and 91.33%. Considering that it is very difficult to match precise bacterial outlines by manual or automated labelling, it unrealistic to anticipate that the form of the predicted contour would precisely match the shape outline of the ground-truth contour. Therefore, even if the model accurately predicted a contour, the errors in the reference used as the ground truth itself may penalise it slightly.

Table 7.1: A comparison of segmentation results between the original UNet and the proposed network. In the training phase, a composite of both stained FOVs was used, whereas in the testing phase, both models were initially evaluated using only green and then both types of FOVs. The LTR dye stained more artefact, making it more difficult to detect *Mtb* cells on the red images precisely. Although the original UNet performed better in training, the proposed network performed better on unseen test data.

Models	Training		Test	
	Dice coefficient	Jaccard index	Dice Coefficient	Jaccard index
UNet (Baseline) Green FOVs	99.53%	99.07%	96.10%	92.49%
UNet (Baseline) All FOVs			91.29%	83.97%
Proposed network Green FOVs	99.04%	98.11%	97.00%	96.06%
Proposed network All FOVs			92.03%	85.24%

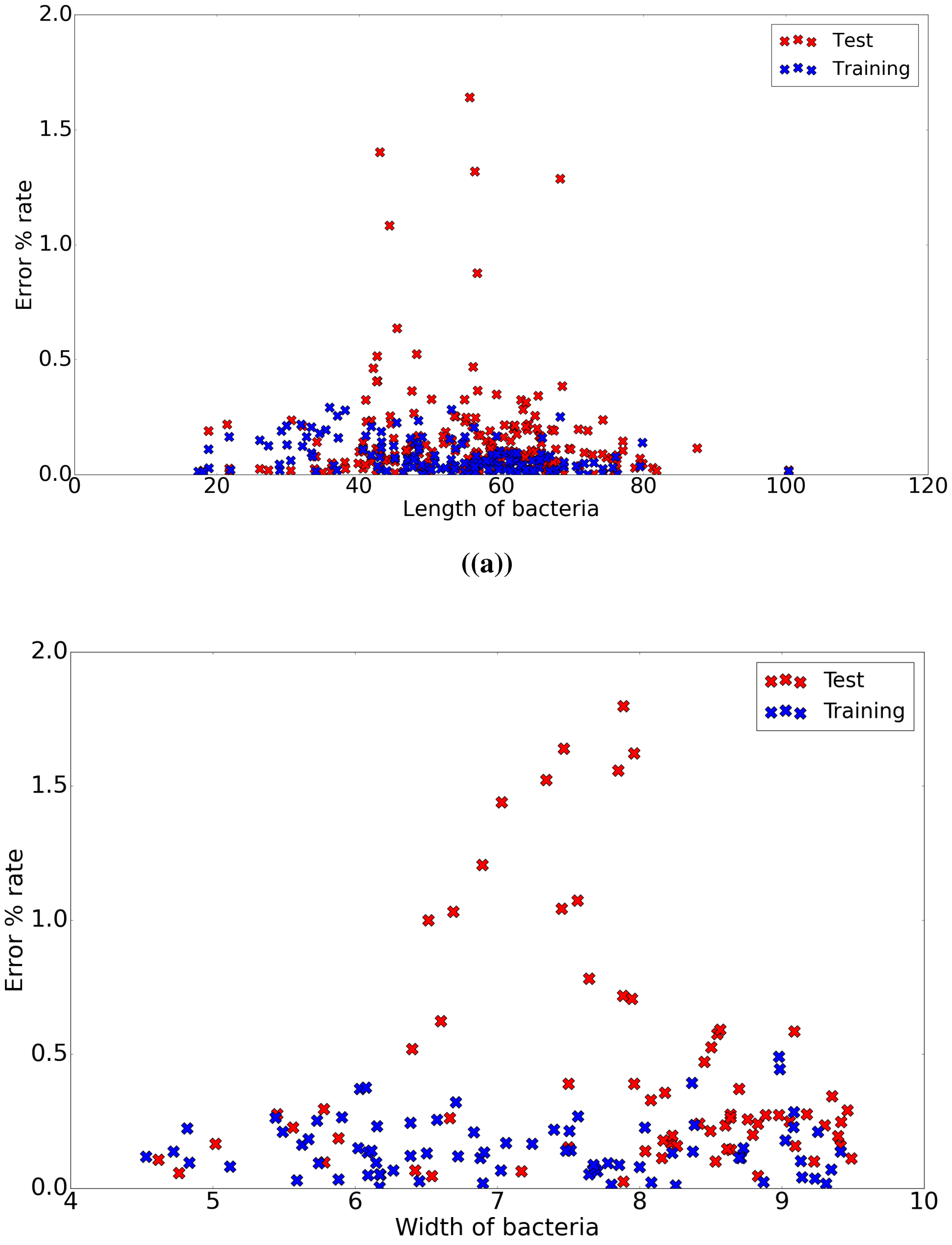
7.4.3 Distance-based evaluation

Having assessed green and red channel image segmentation for all FOVs separately, I evaluated the ability of the network to detect the same bacteria in both images of each FOVs at the same location. I utilise the L_1, L_2, and L_∞ norms in a manner similar to that described in Section 5.5.2.1. Instead of comparing ground truth contours to predicted contours, contours from the green FOV and the red FOV were compared in this chapter. Essentially, I am attempting to correlate the centroids of bacteria in the green FOV with the centroid of bacteria in the red FOV. The pairing was determined by the minimum Euclidean distance between the centroid positions, with a threshold of 15 if an apparent bacterium in one image could not be matched with a partner in the other. If no suitable contour is obtained in the red FOV, the contour from the green FOV is discarded, since it is deemed irrelevant. The combined distances constitute a vector which is subsequently used for the norms calculation. Additionally, I provide the counts of paired contours from each category, namely, green and red ground truth FOVs, along with their corresponding predicted counterparts. The L_1, L_2, and L_∞ norms for the ground truth FOVs measured at 1010.77, 49.17, and 8.54 pixels, respectively, with a total of 572 pairs. Equivalent values for the predicted FOVs, were 1067.7, 56.12, and 9.85 pixels, with 577 pairs. The close proximity of the norms between the two sets of FOV pairings indicates that my technique accurately predicts and pairs images of the same bacteria co-stained with *Mtb* and intracellular lipid detection dyes. Notably, the L_∞ norm, representing the maximum absolute distance, has a difference of less than 2 pixels, and the total of all distances is within 70 pixels of each other. Considering that the average length of a bacterium can range from 20 to 100 pixels, these numbers suggest that the predicted pairings closely align with the ground truth ones.

7.4.4 Bacterial length and width

As described in Section 7.3.3, I use regression to estimate the individual length and average width of bacteria. Therefore, I applied regression evaluation metrics, comprising of RMSE, MAPE, and MAE. The rationale for incorporating both MAPE and MAE lies in the dissimilar nature of the scaling of length and width. As depicted in Figure 7.7, it is evident that the scaling of length ($\approx 20-100$ pixels) and width ($\approx 4-9$ pixels) differs significantly, thereby an error in length would not have an equivalent effect as an error in width. This figure also indicates that MAE is a more suitable loss function than MSE for this dataset because the outliers, represented by the two tails of the distribution, exhibit a smaller deviation from 0% error, while the majority of errors occur on the average samples. This is due to the fact that the outliers, represented by the two tails of the distribution, exhibit a smaller deviation from 0% error, while the majority of errors occur on the average samples.

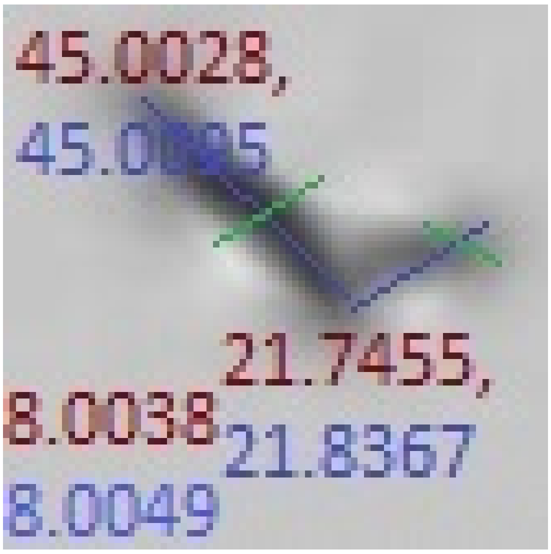
Altogether, all model combinations performed well, with the CNN + HOG combination consistently performing best according to all the criteria. Figure 7.8 shows two examples of cell dimensions measurements using the best model. Table 7.2 summarizes all training and test set metrics. Two additional plots depicted in Figure 7.9 and 7.10, derived from the test set, provide supplementary evidence that the model has exhibited noteworthy performance and has acquired the ability to extrapolate to novel data.



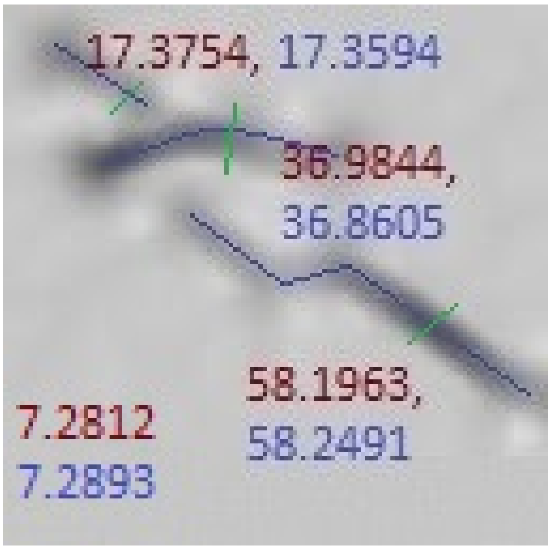
((a))

((b))

Figure 7.7: Plots (**a, b**) length and width samples vs their respective percentage error rate. Due to the large number of samples in both the training set and the test set, the graph is simplified by averaging identical samples with 0.01% error difference.



((a))



((b))

Figure 7.8: Examples of patches to illustrate the labelling procedure for cell dimensions, and show examples of ground truth and prediction distances. The length of bacteria are shown by the blue straight lines while width is depicted in green straight lines. When the length of a curved or angular bacterium requires several blue lines for full coverage, its total length is calculated as the sum of all the blue lines within it. Distances written next to individual cells in blue are ground truth lengths in pixels, while those in red are predicted lengths. The width value is the average of all green lines in each patch and is written in the bottom left corner of each image.

Table 7.2: Performance evaluation metrics for both training and test sets, including all model and shape characteristics. The quantitative results demonstrate that this approach has learnt and generalised the problem. The CNN model utilised throughout the evaluation phase was the pretrained model specifically designed for this purpose.

	Training			Test		
	RMSE	MAPE	MAE	RMSE	MAPE	MAE
CNN	1.9840	0.0213	0.5366	2.4746	0.1111	1.7442
CNN & HOG	**0.0161**	**0.0046**	**0.0212**	**0.0815**	**0.0112**	**0.1004**
CNN & SIFT	0.6727	0.0350	0.6753	0.8357	0.0533	1.0778
CNN & MKD	0.5374	0.0290	0.5469	0.6915	0.0431	0.8732
CNN & HardNet	0.4651	0.0246	0.4646	0.5307	0.0339	0.6628
CNN & HardNet 8	0.7747	0.0393	0.7599	0.9575	0.0615	1.2421
CNN & HyNet	0.5034	0.0263	0.5162	0.6476	0.0402	0.8076
CNN & TFeat	0.1322	0.0077	0.1572	0.1563	0.0096	0.2068
CNN & SOSNet	0.4718	0.0256	0.4948	0.6248	0.0394	0.8007
CNN & LBPs	0.7684	0.0396	0.7624	0.9737	0.0626	1.2495

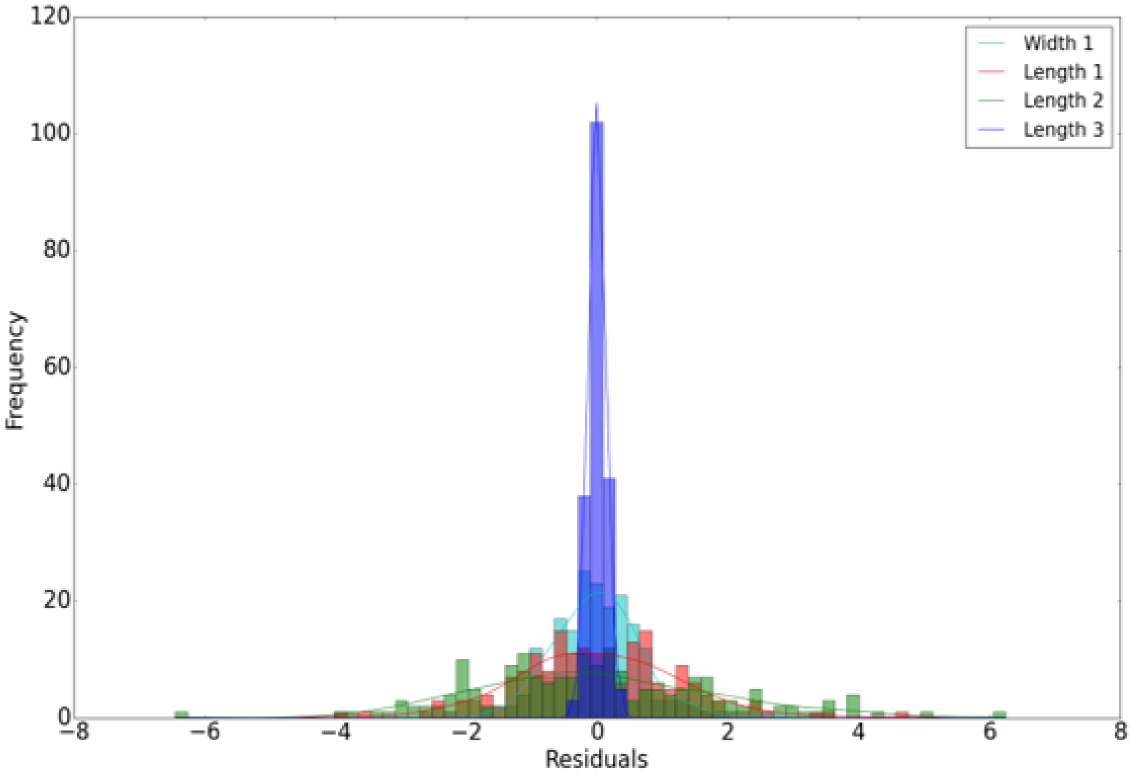


Figure 7.9: The histogram of residuals plot depicts a concentration of residuals around 0, indicating that the model's residuals are predominantly distributed in close proximity to the origin. Patches consisting of 3 or 4 cells are infrequent. As a result, the third length is typically 0, which aligns with the model's accurate prediction. For clarity, the fourth length is not displayed.

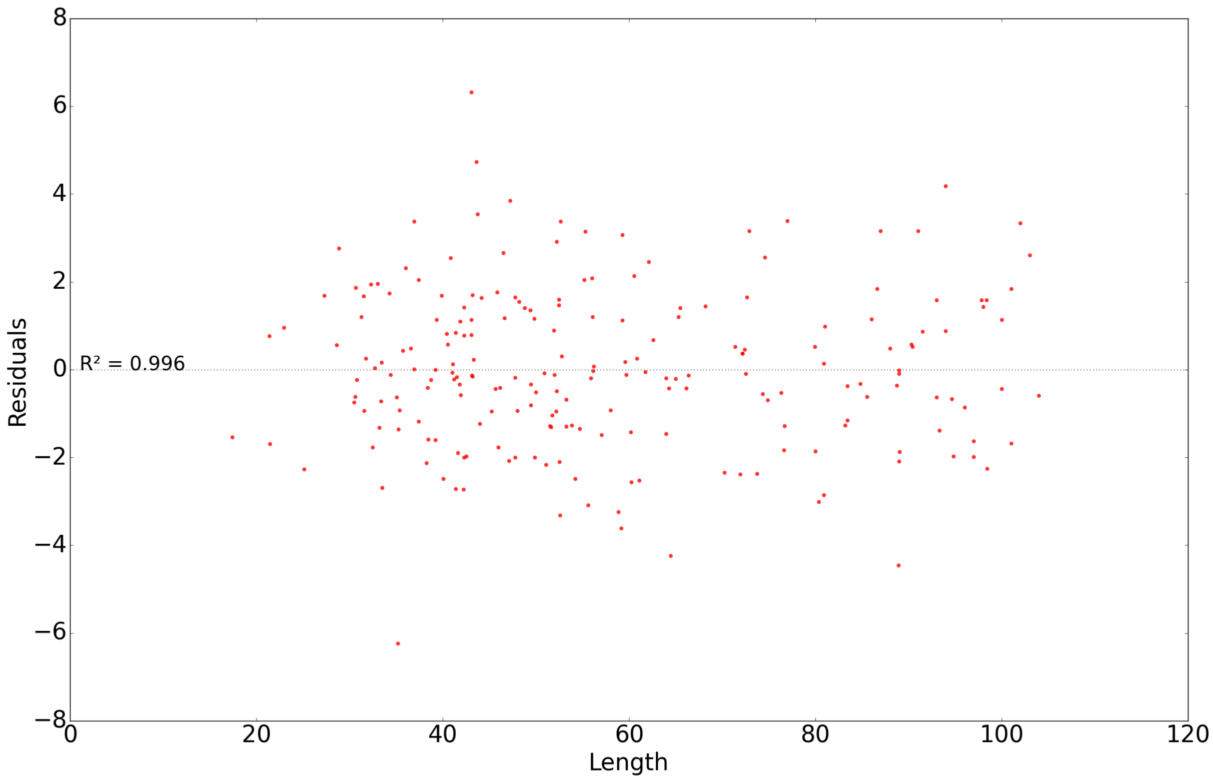


Figure 7.10: Residual plot indicates a dispersion of residuals that is close to zero. An additional observation provides further evidence for the selection of MAE as a more appropriate loss function, given that the outliers in the test set exhibit a proximity to zero that exceeds that of the average sample.

7.5 Discussion

The majority of ML and DL research on automating sputum smear microscopy has focussed on its long-established role as a frontline diagnostic test for pulmonary TB as discussed in Section 3.10. As molecular tools, such as Xpert® MTB/RIF, replace this function, a key contribution of microscopy may become its ability to report on phenotypic characteristics of individual *Mtb* cells for treatment monitoring and to improve our biological understanding of therapeutic response. The work I have descrtibed here is the first demonstration of AI approaches for this application.

I have developed a novel semantic segmentation method for detecting *Mtb* bacteria in fluorescence microscopy FOV which demonstrates superior performance compared to the technique described in Chapter 5 as well as other comparable methods in the field [209, 60, 115, 148]. Additionally, my method is robust for use with multiple fluorescence stains so that paired images of the same FOV can be used to report on bacterial detection and the presence of important intracellular structures such as lipid content. Although the ultimate objective is to estimate the proportion of LR *Mtb* cells within a given FOV, the method described in this chapter does not follow that activity through to provision of a final microbiological result. Presently, the method accurately detects the location of *Mtb* bacteria within dual-stained FOVs and pairs these localized locations based on a specified threshold distance. This is the essential tool for the LR proportion to be determined. Chapter 8 will illustrate the potential of this approach by showing results of pilot experimental work conducted through my AI techniques which demonstrate changing intracellular lipid phenotypes in a sub-population of TB patients during treatment .

Finally, a novel contribution of my work is that the models accurately predicted the dimensions (length and width) of cells in original ground truth images, which does improve the ability of clinical researchers and microbiologists to investigate the relevance of heterogenous bacterial appearances in biological samples. Pilot experimental work in Chapter 8 will also show how my approach could be applied in practice.

The existing limitation of my approach to predicting cell dimensions is its presumption that the number of bacteria in a given patch is unlikely to exceed four. While I do not claim this to be impossible, the probability of such an occurrence is considered (based on observations from both datasets used in this thesis) extremely low. The current design of the technique discussed in this chapter is operates under the assumption of a maximum of four bacteria and may result in either an excessively high length estimation for one bacterium or an attempt to distribute uniformly along the length of all bacteria. In either case, this may lead to inaccuracies in the results.

Next steps for this work will include: i) interdisciplinary collaboration between Infectious Disease and Computer Science researchers to deploy these tools on bigger microscopy image sets to assess their real-world application, and ii) optimisation of methods for automated reading of whole slides, so that the manual labour required to identify FOVs and patches before DL techniques can be used is also eliminated; in effect this work would combine the progress described Chapter 6 and 7.

Overall, the information compiled in this chapter has shown that microscopy based treatment monitoring and *Mtb* cell phenotyping research is important, and that automated DL techniques make these tasks possible.

Chapter Eight

General Discussion

Chapter Abstract – The concluding chapter of this thesis offers an overall summary of the work. Initially, I reflect on the most important results and contributions of my research, explaining in detail how each element contributed to advance the understanding of some major questions identified in Chapter 3. In addition, I describe the limits of my work and the ways in which they may be improved. In conclusion, I provide proposals for future research in the field.

8.1 Chapter introduction

Microscopy dates back to the 17th century, when Dutch scientist Antonie Von Leeuwenhoek discovered microbial life forms using a rudimentary light microscope, a discovery that would spark a revolution in science [251]. Oskar Heimstaedt built and used the first functional fluorescent microscope in 1911. Although fluorescence microscopy was pioneered during the last century, it remains widely deployed in clinical microbiology and has been the mainstay of pulmonary TB diagnosis until recently. However, several centres throughout the world are now shifting their focus away from smear microscopy and toward molecular tools (such as the Xpert® MTB/RIF test) for TB diagnosis [140]. This has led some scientists to question the research and financial cost of automating an older technology like microscopy with tools from the past decade, DL. As described in Chapters 1 and 2, the largest burden and majority of TB-AI research occurs in LMICs, and the expense of deploying expensive computer-based technologies, and training personnel to use AI methods may prove problematic. On the other hand, once deployed AI methods can be extremely labour-saving and there are some potential applications of fluorescence microscopy (particularly in the field of TB treatment monitoring and therapeutics research) which cannot be easily replaced by other methods. Whilst this thesis does not argue for wholesale resurgence of microscopy as the primary diagnostic tool for TB worldwide, I advocate that it still has an important and irreplaceable function. Since microscopic examination of *Mtb* cells remains critical as a research tool, work on the development of improved automated tools for standardising and expediting image analysis remains important.

From that starting point, the research described in this thesis has sought to determine whether AI approaches may aid detection, quantification and characterisation of *Mtb* cells from fluorescence microscopy images. Chapter 3 reviewed progress in this field prior to my work, and highlighted some unanswered questions. Chapters 4 to 7 have described original research to advance methods to fill some of the identified gaps. In this final chapter I will:

- Review essential findings and their contributions to the research field from each successive chapter, restating the most significant outcomes of my work.
- Consider future directions, which could build on the work I have done so far, including limitations of my TB-AI approaches and ongoing challenges.
- Conclude by outlining my proposals for future action.

8.2 Chapter contributions

Before delving into detailed discussion of each chapter contribution, I will first present a summary of the key points that represent the primary contributions of this thesis. While these points are not exhaustive, they encapsulate the core achievements:

- A technique for image enhancement designed to standardize datasets, enhance the visibility of bacteria, and support the learning process for detecting *Mtb* cells across various staining methods.
- A novel architecture created to exclude FOVs devoid of bacteria, applied subsequent to slide segmentation.
- A method that sets a new standard for accuracy in detecting *Mtb* cells using dual stains within a single training cycle, where the use of dual-stained FOVs enables the estimation of lipid content.
- The introduction of a unique architecture paired with feature descriptors serving as feature extractors, which are then utilized by a MSVR to estimate the dimensions of cell length and width.

In Chapter 3, I reviewed efforts to automate analysis of sputum smear microscopy images, which have gradually advanced over a period of more than twenty years. Progress has been made but several obstacles remain to be addressed. A significant limitation is the absence of comparative analyses between the different TB-AI methodologies that have been described. Most research groups work on their own proprietary image-sets, which vary because of qualitative differences in sample preparation, microscopy protocols and imaging techniques. These image sets are rarely shared on-line alongside published manuscripts. I curated those which are currently publicly available, encompassing relevant extracted data along with any supplementary annotations. The absence of standardisation in existing datasets is important because the influence of the data used on the reported efficacy of methods has been observed to be substantial. Methods which work well on one set of images (often the set that CV algorithms have been designed with, or ML/DL methods have been trained on) often perform less well on unseen data. If researchers were willing to share the image sets which they used in open-access online repositories it would be possible for groups to develop techniques on one image set, then evaluate them on alternative independent data to assess generalisability. This may also help address the ever-present danger of model-overfitting.

On a related theme, it would be beneficial to standardise benchmarks for evaluation of each category of TB-AI microscopy activity (classification, regression, and segmentation), so that work done by different researchers can be compared using the same metrics, even if those researchers also choose to employ their own additional metrics. Encouraging diverse research teams in different settings worldwide to think and act in similar ways is challenging. However, there is precedent for development of a ‘minimum standards’ approach to presenting and analysing data in other aspects of infectious research; including pharmacogenetic studies [35], and outcome reporting for clinical trials of new TB treatments (included within the Core Outcome Measures in Effectiveness Trials initiative, `https://www.comet-initiative.org`) [26, 28, 27]. A similar strategy could be advocated here.

Notwithstanding these challenges, ML and DL techniques within the TB-AI microscopy field have achieved notable successes and each approach possesses its own strengths. It has been demonstrated that the detection of *Mtb* bacteria necessitates reliance on pixel intensity and shape, regardless of whether the approach employs ML or DL techniques. Most prior work focusses purely on identifying *Mtb* cells, rather than counting them or describing their detailed phenotypes. Most image sets used for existing work comprise pre-selected FOVs rather than WSIs, meaning that manual microscopy work is still required to crop the slide into manageable images. My work has investigated under-studied topics by tackling bacterial quantification in Chapter 5, WSI analysis in Chapter 6, and cell phenotyping in Chapter 7.

One limitation of the work presented in Chapter 3 pertains to its exclusive focus on microscopy. Other contributions in the field of TB-AI that involve the analysis of CT scans or chest radiographs from patients with pulmonary TB were excluded. Some of these approaches may have potentially made valuable contributions to the analysis of microscopy images. For instance, a method utilised for CT scans could potentially be adapted and applied to microscopy images, albeit with modifications to account for the unique characteristics of different image types.

In Chapter 4, I explained that development of ML/DL methods for TB-AI microscopy require large carefully annotated training sets of images from the outset. Re-deployment of clinical and microbiology staff from the School of Medicine to front-line Covid-19 pandemic response duties during the early period of this PhD meant that image annotation was delayed, so I took the opportunity to explore the viability of *Mtb* detection solely using conventional CV methods. Specifically, investigation was undertaken of whether semantic segmentation is achievable utilising a FOV, a ridge-based detector, and geometry-based features. Without the automated learning component of ML/DL, this approach required input of heuristic knowledge of bacterial morphology by the model developer. When a CNN slides across an image to execute convolution between the image and its filters, for instance, the filter weights (also known as trainable parameters) are dynamically adjusted to learn the task at hand. The same cannot be true for image derivatives, such as the Gaussian filter used to generate image derivatives in CV work. When executing through the full dataset, the weights of such filters remain unchanged. Although the typical rod-like morphology of *Mtb* is consistently defined, many practical examples demonstrate that these cells may adopt a large number of, often unpredictable, shapes. The concept that employing a ridge-based detector, computing the second-order derivatives of an image using the Hessian matrix and a Gaussian filter, and then analysing its eigenvalues will provide satisfactory detection of all bacilli was refuted. Detection of bacteria is not easily accessible by only obtaining the image derivatives, which are dependent on pixel intensities, i.e. colour, as was shown in the literature and confirmed in Chapter 4. Although the findings of this chapter indicated that CV algorithms alone were insufficient for bacillary detection, an image enhancement technique based. lower value Hessian eigenvalues when images are converted to greyscale was seen to be beneficial in separating signal from noise and was carried forward to the work of subsequent chapters.

Chapter 5 sought to address the question,“can an automated image analysis pipeline provide an accurate estimate of bacterial load, incorporating all *Mtb* bacillary morphologies?” An ablation study to evaluate the method developed in this chapter revealed that each pipeline module performed relatively well for its assigned purpose and contributed to strong overall performance. The initial step of the pipeline, segmentation, was comparable to, but did not improve upon, other efforts of the same kind. This step was created with a higher emphasis on sensitivity than specificity in order than all possible objects of interest in each FOV image advanced to the next stage, or else they'd be excluded from further analysis (running the risk that some true positive bacteria would be overlooked). The classification step was then responsible for excluding false positive objects, improving the pipeline's specificity. However, one of the drawbacks of the segmentation step is that Cycle-GANs often fail to generate symmetrically rectangular bounding boxes around bacteria. This effect necessitates additional processing prior to classification or overlap metrics such as the Jaccard index are detrimentally affected. The final stage of the pipeline, autonomous bacterial quantification by regression, has acquired a new state of the art in the field. A key finding of this chapter is that generic models are too complex for this task; thus, for best results, it is essential to create task-specific model architecture. Moreover, each phase of the pipeline may function as a standalone component so long as its input requirements are satisfied.

Attempts to detect and quantify *Mtb* cells in Chapters 4 and 5 still start from the manually pre-selected FOVs of Image Set 1. The process of generating these FOVs remains laborious and subjective so it is essential to automate this step if TB-AI microscopy is really to advance as a treatment monitoring and research tool. Creation of new automated methods for standardising and accelerating whole slide image processing in order to remove human participation from FOV selection is the objective of Chapter 6. In this chapter, I described a novel solution based on a newly designed DL-based architecture tailored specifically for the task, which learns from coarsely labelled FOV images and the corresponding binary masks, and then classifies novel FOV images as containing or not containing bacteria. Contrary to Chapter 5, the method in this chapter was initially focussed on achieving high specificity so that no false positive FOVs advanced to future evaluation. Initial findings were promising, since my suggested model surpassed all generic models on accuracy, and detected no FPs at all, the model's inability to account for all true positive objects might be cause for caution. As described in Chapter 6, the sensitivity and specificity parameters of my model can be amended, according to desired output, by tuning the decision threshold to minimise FNs whilst increasing FPs. The cost of the trade-off to do this with my model was lower than that of the second best model, emphasising that the proposed model is more adaptable than others.

Finally, Chapter 7 examines the viability of TB-AI microscopy in relation to detection of *Mtb* cell phenotypes which might influence response to therapy and may change during treatment. From my literature review in Chapter 3, no one has undertaken such a work previously, i.e. automated treatment monitoring inference by the use of microscopy for TB. Using a custom-built CNN as a pre-trained encoder in combination with feature descriptors, I was able to obtain highly promising results in the area of semantic segmentation for *Mtb* bacteria and intracellular lipid detection, and cell length/width estimation. An important contribution of this chapter is that it enables microbiology researchers to investigate the impact of drug-tolerant *Mtb* bacteria further, they may do so without spending excessive hours on subjective cell-by-cell microscopy. The proposed method has advanced the state of the art in *Mtb* bacteria semantic segmentation in compared to prior publications. In addition, it is capable of producing very accurate estimates of cell length in both training and test sets, highlighting the generalisability of the model. The proposed method may also be used for bacterial load assessment, as it may indirectly infer the number of bacteria present in the patch based on the number of lengths calculated, because each cell has one length unit. Comparison of the performance of this model with other methods to estimate cell dimensions is not possible since no other work in TB treatment monitoring has attempted that specific task. One of the limitations of this chapter is the innovative use of FOVs for Auramine O and LTR stains within the same dataset. Employing LTR stained FOVs needs more improvement in appropriately detecting *Mtb* bacteria, since both evaluation metrics were adversely affected by the use of these FOVs.

8.3 Future work

I conclude the thesis with a summary of the components of this topic that I believe should be take into consideration for future research. Before moving further with future work, it is necessary to emphasise where microscopy is presently most relevant to TB-AI research.

Based on the presented data and findings, I propose the following topics for extending research in TB-AI microscopy:

- Improving image segmentation for bacterial detection using crowd counting or Vision Transformers.

- Improving the accuracy of bacterial load estimation by progressively minimizing the percentage error through the use of density maps.

- Improving the computational efficiency of FOV acquisition by employing a hybrid method that combines clustering algorithms with shape-based classification to isolate salient FOVs.

8.3.1 Improving *Mtb* bacterial detection

Improving the accuracy of bacterial detection is essential, as the efficacy of all subsequent downstream research activities depends on the reliable detection of bacteria to realize the full potential of automated methods and integrated pipelines. As an illustration, it becomes evident that optimizing the segmentation stage within the Chapter 5 pipeline could potentially yield improvements in the performance of both classification and regression stages. To improve achievement of this aforementioned task, I suggest the following methodological approaches:

- *Mtb* bacterial detection with crowd counting.
- Vision transformers for semantic segmentation.

8.3.1.1 *Mtb* bacterial detection with crowd counting

Due to its practical use in surveillance systems, crowd counting is an important problem in CV. The standard architecture of crowd counting algorithms consists of two phases, namely global regression and density estimation [246]. The latter approach, which involves predicting a density map that is subsequently summed to obtain the final count, typically exhibits superior performance compared to the former, which estimates the final count directly from the image. This advantage arises because the latter method leverages additional spatial information through the density map. Although crowd counting approaches were created with crowd surveillance in mind, they may be used to any task involving densely populated, homogeneous objects of interest in an image [249, 245, 124, 36, 181]. For the labels of this task, it is essential to generate ground-truth density maps based on crowd images (density map synthesis). In these maps, the object of interest can be represented as specific image locus [124] which serve as focal points for object counting or segmentation. Density map estimation is the process of designing DL models to predict a density map from an input image. Most research efforts have concentrated on the task of density map estimation [246]. However, such a task (as with most ML/DL problems) is limited by the creation of annotated datasets, in this instance the compilation of adequate and accurate density maps. The hand-crafted techniques employed for generating density maps may not be ideally suited for end-to-end training with the specific network or dataset in use. Indeed, in this scenario, it must be determined whether the image locus represents a single bacterium. In addition, it must be determined if the crowd count will be used for segmentation or to estimate the number of bacteria. The former may accept a single image locus for several bacteria that overlap (i.e. bacterial contiguous location), but the latter will presume one locus per bacterium, even in the case of clumped bacteria. In this instance, if the crowd counting approach will only be used to segment the image into tiny patches based on regions of interest, then the number of clumped bacteria is immaterial, since they will be handed down for further processing, i.e. a model appropriate to that task. In addition to the density map, the training component of this method needs a standard encoder–decoder model, such as the convolutional layers of the VGG as the encoder as was done in some works [130, 203].

8.3.1.2 Vision transformers for semantic segmentation

Despite the widespread adoption of the Transformer architecture for natural language processing (NLP) applications, its utilization in CV is rapidly expanding. Rather than being confined to roles where it works in conjunction with CNNs or replaces specific components of CNNs, Transformers now serve as the core architecture for various standalone CV tasks [66]. Considering that semantic segmentation is a subfield of vision, it has the possibility to employ a Vision Transformer (ViT) to generate a binary masked image in which bacteria are labelled in white and everything else is labelled in black. In contrast to convolution-based approaches, this methodology enables modelling of global context at the first layer and throughout the network, while in a CNN this occurs gradually and intermediately as the image traverses the network. Strudel *et al.* [215]employ the recently introduced ViT and extends it for semantic segmentation by relying on the output embeddings corresponding to image patches and obtaining class labels from these embeddings using a point-wise linear decoder or a mask transformer decoder [37]. The extensions of a linear decoder or a mask transformer decoder hence enables the ViT to upsample (rebuild) a binary version of the input image.

ViTs typically necessitate a higher number of trainable parameters compared to CNNs due to the inherent complexity of their architecture. While CNNs employ convolutional, pooling, and fully connected layers to efficiently share parameters across the image, ViTs rely on self-attention mechanisms for feature extraction. The flexibility offered by self-attention comes at the expense of additional parameters [66, 232]. Furthermore, ViTs often require higher resolution inputs than CNNs to effectively capture fine-grained information, contributing to an increase in computational requirements rather than the number of parameters. Although ViTs excel in various CV tasks, especially with large datasets, they are not well-suited for tasks involving small patches, as discussed in Chapters 5 and 7.

Regarding training time, ViTs generally require longer durations compared to CNNs due to their computational complexity. The self-attention mechanism involves numerous computations across the entire image, which can be both memory-intensive and time-consuming. Techniques such as knowledge-distillation (KD) [99] or pre-training on extensive datasets like ImageNet can mitigate this by reducing the number of parameters and thereby improving training efficiency. Overall, the training time for a ViT is influenced by multiple factors, including its architecture, input image size, and the implementation of optimization techniques such as knowledge-distillation (KD) or pre-training. Overall, the training time for a ViT can depend on a variety of factors, including the specific architecture, input image size, and the availability of pre-trained models or the use of distillation techniques.

While it may require more training time than a CNN, the potential benefits of using a ViT, such as improved performance on certain tasks, may justify the additional computational cost. Sagar proposed a hybrid ViT architecture for biomedical image segmentation that resembles a UNet architecture (described in Chapter 7) because its architecture has a U shape [193]. The input image is divided 16×16 pixel patches which are then fed into the embedding layer as well three 1×1, 3×3, and 5×5 feature maps are generated for each patch. After they have been concatenated and vectorized, they are fed into 3 encoder transformer blocks that lead to 3 decoder transformer blocks with UNet-like skip connections before being linearly projected to produce the segmented image [193].

8.3.2 Improving bacterial load estimation

As initially described in Chapter 3, the literature on bacterial load estimation of *Mtb* is very limited, with three out of five publications offering an end-to-end automated approach.

8.3.2.1 Cell counting with the use of density maps

As stated in Section 8.3.1.1, density maps may also be used for cell counting, a technique used to estimate the number of cells in a microscope image. To generate a density map, the image must first be segmented so that individual cells can be detected. This is possible using the techniques presented in Chapters 5 and 7. Other potential solutions may be Fast and Faster R-CNNs [183, 85].

After detecting the cells, their centroids are computed and indicated on the image. *Mtb* bacteria density maps may still need human annotation due to the possibility of clumped bacteria, even after automated segmentation. A grid of square or circular sections is then superimposed on the image. The zones should be small enough to correctly record the cell density, yet big enough to reduce the influence of noise [253]. The number of cells in each area can then determined by measuring the intensity of the pixels in that region's density map. The total number of cells in the image may then be calculated by adding the number of cells in each grid area and multiplying by the inverse of the proportion of the image covered by the grid [253]. Density map cell counting is a potent and extensively used technique for estimating cell density in microscopy images. It may be used to compare the density of cells in various sections of an image or between images. However, this approach may not be very accurate when bacteria are clustered or when the image includes noise (artefacts). One approach for mitigating this problem involves dataset augmentation through the generation of synthetic images by overlaying outliers from the dataset, with the objective of producing new input and corresponding density maps.

8.3.2.2 Task–specific model architecture

As shown in Chapters 6 and 7, generic models are often too complex for the datasets of this project, as my proposed methods of these chapters performed better than the generic ones under identical environmental setup and conditions. Specifically, despite being augmented, the dataset is still substantially smaller than the number of parameters of the generic models used, as observed by the results in Chapter 5. Moreover, the combination of the positive outcome presented in Chapters 6 and 7 and the negative findings presented in Chapter 4 provides additional confirmation that *Mtb* cell detection favours a two-step approach, i.e. object selection based on both appearance and shape characterisation. Thus, the evidence supports the promising potential of custom-designed architectures for bacterial load assessment.

8.3.3 Improving computational efficiency of FOV acquisition

As mentioned, the most useful function of sputum smear microscopy is not to simply classify entire slides as ‘TB positive’ or ‘TB negative’, but to extract relevant FOVs for more detailed subsequent analysis (e.g. counting the number of cells in them or describing those individual cells in detail). Clinical samples, particularly of non-sterile specimens like sputum often comprise complex mixtures of cell populations, and identifying and characterizing microorganisms can prove difficult, particularly when they are sparsely distributed [86]. Focusing on specific regions of interest allows researchers to direct their attention towards microorganisms of particular relevance for their investigations. Moreover, this practice reduces the volume of data to be processed and the computational burden of analysing extensive and intricate datasets [61, 99]. The ideal methodology for TB-AI research would directly take a microscopy slide as input and produce comprehensive outputs, including cell count, phenotypic characteristics, and lipid content. However, currently, no such all-encompassing method exists in the field.

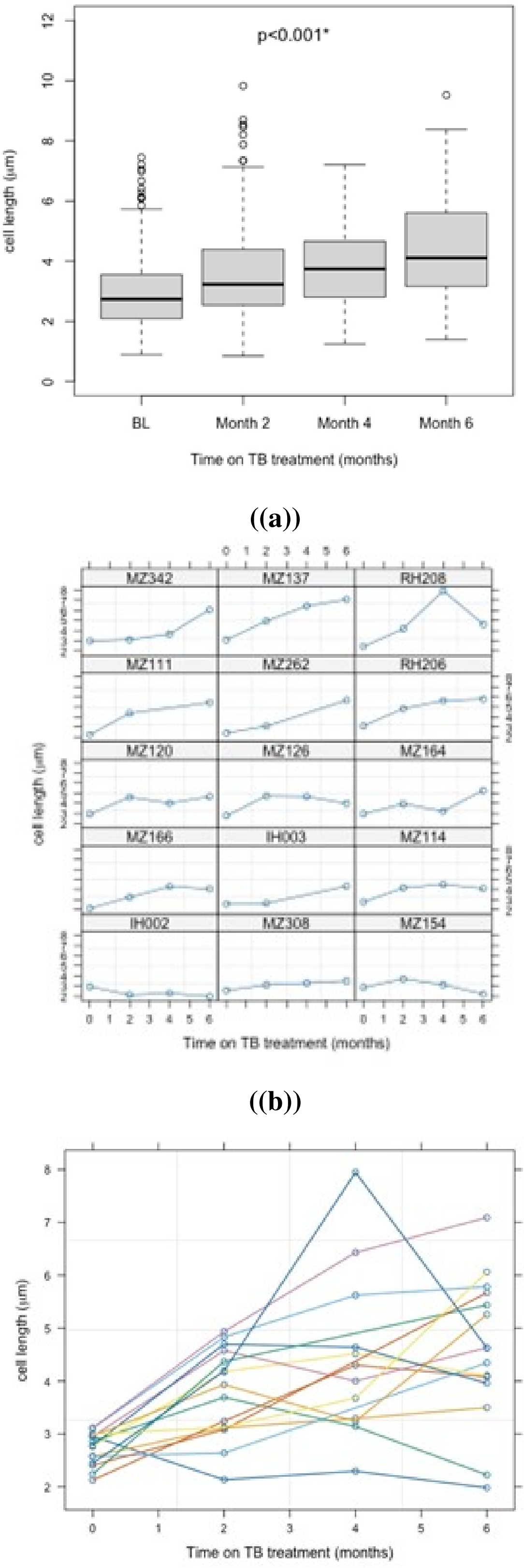
Therefore, evidence indicates that for effective downstream TB-AI analysis, it is necessary (thus far) to crop and magnify the slides. As seen from existing literature, the prevalent approaches to achieve this involve either employing auto-focus algorithms or resorting to sequential cropping methods. Although the latter has a lower risk of false negatives, it is computationally demanding since the entire slide must be cropped. To minimise superfluous cropping and magnification, a hybrid approach that combines these two techniques could be adopted. Each tile of the slide is loaded and pre-processed with image filters essentially creating a ridge/edge-based detector similar to that described in Chapter 4. Since the tile is now binarised, clustering techniques can be applied to classify the latent distribution based on shape characterisation as shown in Chapter 6.

8.4 Potential application of new methods for treatment monitoring and detection of bacterial phenotypes

Arguably the most compelling arguments for ongoing sputum smear microscopy lies in its capacity to monitor the progress of TB treatment and to report on how the composition of *Mtb* cell phenotypes changes during antibiotic exposure. My work on automation of these activities are central to the conclusions drawn from this thesis, and to the potential direction of future work. To illustrate how my methods may be applied, I performed pilot work on sputum smear microscopy images collected serially from the first 15 patients in Image Set 1 at 0, 2, 4 and 6 months. Even for 15 patients, these experiments required detailed evaluation of 914 bacteria. This work would require a prohibitive amount of time to conduct manually at larger scale.

The simple hypothesis to be tested is that *Mtb* cells increase in length and lipid content as TB treatment progresses. To reprise and expand the background relevance of this hypothesis, recall from Section 2.7 that mycobacteria have unusual growth characteristics. They do not divide symmetrically with a division septum at the centre of their cells. and they elongate unevenly by adding peptidoglycans at their poles [39, 7, 226]. Longer cells may be more rifampicin tolerant [184] and associated with more severe clinical disease [243], and the bacilli seen at microscopy may incrementally lengthen after antibiotic exposure [19]. Additionally, LR *Mtb* cells may be more antibiotic tolerant, and their emergence could be associated with worse outcomes in pulmonary TB treatment [207, 96, 57].

Using my approaches from Chapter 7, the Infectious Diseases Group in the School of Medicine and Orange Jellyfish Computer Science Laboratory at the University of St Andrews charted changes in cell length and the proportion of LR cells in over time for 15 Tanzanian patients (see Figures 8.1 and 8.2 respectively). The median cell length increased from 2.74 to 4.11μm between 0 and 6 months and fitting a linear mixed effects model to available data from these patients suggests an increase in cell length in *Mtb* cells visible at smear microscopy of 0.3μm per month of treatment. To achieve full automation of the lipid content estimation method described in Chapter 7, the procedure commences with the detection of bacilli in the Auramine O FOV. If a corresponding bacilli is identified in the LTR FOV, a green centroid is marked on the Auramine O FOV. Conversely, if a match is not found, a red centroid is annotated on the detected bacillus in the Auramine O FOV. The aggregate of green centroids represents the count of LR cells, while the total of red centroids corresponds to the count of LP cells. The median proportion of LR cells increased from 0.67 to 1.00 between 0 and 6 months, and fitting a linear mixed effects model to the available data suggests an increase in LR cell proportion of 0.05 per month of treatment. Within these general summary statistics, considerable heterogeneity can be seen in the phenotypes of *Mtb* cells observed in different patients at baseline and in trajectories over time. Overall, therefore these preliminary data appear to support the microbiological hypothesis being tested: more importantly, they show that the tools developed in this thesis could be used to test it.

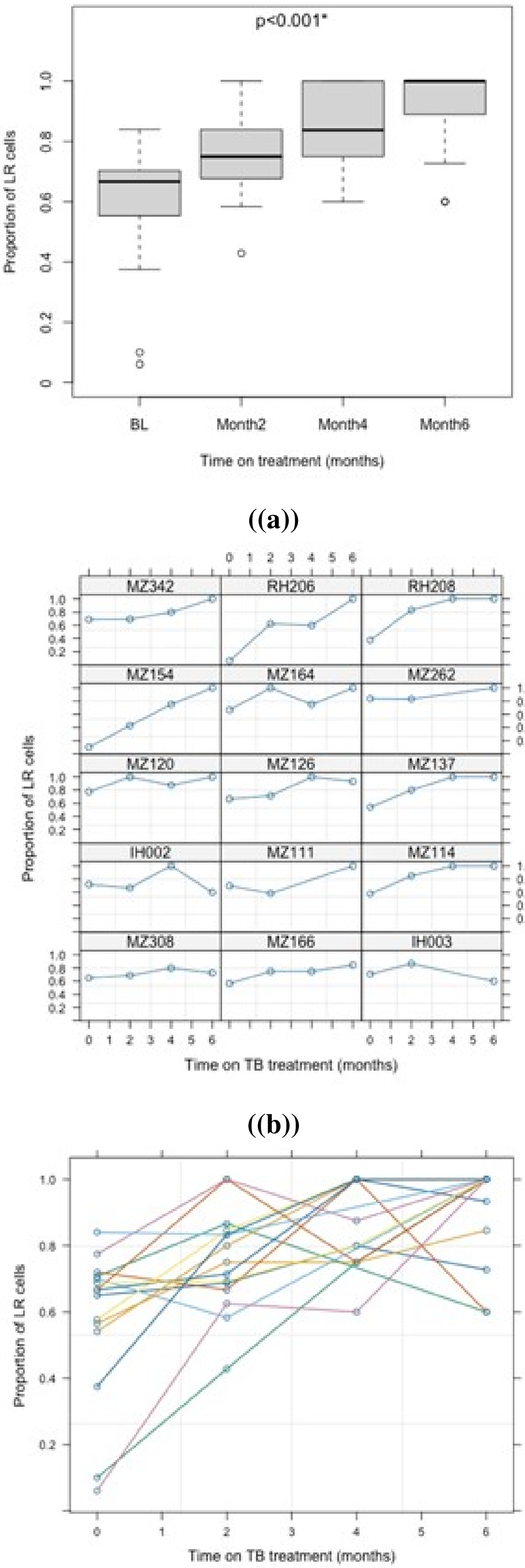


((a))

((b))

((c))

Figure 8.1: (a): Boxplot of cell length at 0, 2, 4 and 6 months of treatment. The DL model estimated cell length in pixels, which is converted to μm by dividing by 15. A statistically significant difference in cell length between the time-points is illustrated by ther Kruskall-Wallis test. (b): Individual plots of changes in cell length for each of 15 patients studies. (c): A spaghetti plot of individual patient data super-imposed on one another



((a))

((b))

((c))

Figure 8.2: (a) Boxplot of LR *Mtb* cell proportion at 0, 2, 4 and 6 months of treatment. A statistically significant difference in cell length between the time-points is illustrated by ther Kruskall-Wallis test. (b): Individual plots of changes in LR *Mtb* cell proportion for each of 15 patients studies. (c): A spaghetti plot of individual patient data super-imposed on one another.

These data should be viewed with caution; they are preliminary observations from a small number of patients and in-depth analysis of them would be inappropriate. Additionally, the sampling timepoints of 0, 2, 4 and 6 months may need to be refined for future studies as very different numbers of cells are visible for assessment at such widely spaced intervals (~40 cells/patient at month 0 but <10 cells/patient by month 6). Nevertheless, my objective here is not to present data or results, but simply to showcase in principle the type of work that could be accomplisjed in a shorter period of time using automated rather than manual methods.

Further analysis using these tools, on bigger datasets could interrogate patient or microbiological factors associated with variable phenotypes at baseline and the importance of changes over time, e.g. does the rate of elongation in cells at microscopy of an individual patient, or does the rate of increase in LR proportion in early TB treatment really have any impact on outcome. Limitations of time prevented such work in my thesis, and I would not seek to undertake it independently in any case as I would need to collaborate with clinicians and microbiologists with deeper understanding of TB. However, the design and conduct of studies to answer these questions would be a natural progression from my work.

8.4.1 Simplifying automated microscopy for clinicians and microbiologists

The work of this thesis has described the computer science basis of DL tools for TB-AI microscopy, particularly in relation of knowledge gaps which were identified in Chapter 3. However, the downstream application of these tools will ultimately be for clinicians and microbiologists rather than computer scientists. For my methods to be implementable by those who need to use them, further engineering of a pipeline approach will be necessary. Ideally, sputum smears could be prepared and digitally imaged on a slide scanner to provide raw data for streamlined automation of classification of FOVs of interest, then regression analysis of cell length and segmentation of individual bacilli for LR proportion analysis. This would require additional refinement of my tools, and software engineering which is considerably beyond the scope of my thesis could be an ultimate outcome of the research. Making such computationally intense tools available for use in the LMICs where TB is high prevalence would be a further enormous challenge, but even developing methods which could be applied to sputum smears transported to translational research laboratories would be valuable.

8.5 Conclusion

Overall, the process of TB diagnosis from sputum samples is changing worldwide, with less reliance on microscopy in many centres and increasing focus on rapid molecular tools such as

Xpert® MTB/RIF. In the medical context, microscopy, whether brightfield or fluorescence, is unlikely to remain the standard-of -care tool for tuberculosis diagnosis in most settings [25, 238, 173]. However, smear microscopy still plays an important role in assessing disease severity and monitoring therapy, and there is considerable precedent and research interest in microscopy-based tools for single cell phenotyping to better understand TB treatment response. Therefore, AI work to automate smear microscopy image analysis remains of value.

My thesis has illustrated key gaps in our existing knowledge, and in our approach to research in this field. I have also advanced the field by showing that original TB-AI microscopy methods, particularly using custom-designed DL models can be applied to: count *Mtb* cells on fluorescence microscopy FOVs, select FOVs of interest from WSIs and estimate cell length and lipid content on appropriately stained clinical samples. If progressed further by interdisciplinary teams of clinicians, microbiologists and computer scientists TB-AI microscopy could still contribute meaningfully to the ongoing global fight to control TB.

References

[1] Example of a line, an edge and a ridge. *Last accessed: 02-11-2023*. URL: `https://i.stack.imgur.com/d6TZy.png`.

[2] R. E. Aarnoutse, G. S. Kibiki, K. Reither, H. H. Semvua, F. Haraka, C. M. Mtabho, S. G. Mpagama, J. van den Boogaard, I. M. Sumari-de Boer, and C. Magis-Escurra. Pharmacokinetics, tolerability, and bacteriological response of rifampin administered at 600, 900, and 1,200 milligrams daily in patients with pulmonary tuberculosis. *Antimicrobial Agents and Chemotherapy*, 61(11):10.1128, 2017.

[3] W. H. Abir, M. F. Uddin, F. R. Khanam, T. Tazin, M. M. Khan, M. Masud, and S. Aljahdali. Explainable AI in diagnosing and anticipating leukemia using transfer learning method. *Computational Intelligence and Neuroscience*, 2022:5140148, 2022.

[4] T. Ahmed, F. Wahid, and J. Hasan. Combining deep convolutional neural network with support vector machine to classify microscopic bacteria images. In *International Conference on Electrical, Computer and Communication Engineering*, pages 1–5, 2019.

[5] H. Albert, Y. Manabe, G. Lukyamuzi, P. Ademun, S. Mukkada, B. Nyesiga, M. Joloba, C. N. Paramasivan, and M. D. Perkins. Performance of three LED-based fluorescence microscopy systems for detection of tuberculosis in Uganda. *PLoS One*, 5(12):e15206, 2010.

[6] H. Albert, R. R. Nathavitharana, C. Isaacs, M. Pai, C. M. Denkinger, and C. C. Boehme. Development, roll-out and impact of Xpert MTB/RIF for tuberculosis: what lessons have we learnt and how can we do better? *European Respiratory Journal*, 48(2):516–525, 2016.

[7] B. B. Aldridge, M. Fernandez-Suarez, D. Heller, V. Ambravaneswaran, D. Irimia, M. Toner, and S. M. Fortune. Asymmetry and Aging of Mycobacterial Cells Lead to Variable Growth and Antibiotic Susceptibility. *Science*, 335(6064):100–104, 2012.

[8] O. Arandjelović. Hallucinating optimal high-dimensional subspaces. *Pattern Recognition*, 47(8):2662–2672, 2014.

[9] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Conference on Computer Vision and Pattern Recognition*, pages 2911–2918, 2012.

[10] G. W Ashdown, M. Dimon, M. Fan, F. Sánchez-Román Terán, K. Witmer, D. C. A. Gaboriau, Z. Armstrong, D. M. Ando, and J. Baum. A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens. *Science Advances*, 6(39):eaba9338, 2020.

[11] S. Ayas and M. Ekinci. Random forest-based tuberculosis bacteria classification in images of ZN-stained sputum smear samples. *Signal, Image and Video Processing*, 8(1):49–61, 2014.

[12] V. Ayma, R. De Lamare, and B. Castañeda. An adaptive filtering approach for segmentation of tuberculosis bacteria in Ziehl-Neelsen sputum stained images. In *Latin America Congress on Computational Intelligence*, pages 1–5, 2015.

[13] K. S. Babiarz, S. C. Suen, and J. D. Goldhaber-Fiebert. Tuberculosis treatment discontinuation and symptom persistence: An observational study of Bihar, India's public care system covering >100,000,000 inhabitants. *BMC Public Health*, 212:418, 2014.

[14] S. H. Baek, A. H. Li, and C. M. Sassetti. Metabolic Regulation of Mycobacterial Growth and Antibiotic Sensitivity. *PLoS Biology*, 9(5):e1001065, 2011.

[15] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *British Machine Vision Conference*, pages 1–12, 2016.

[16] Y. Bao, T. Xiong, and Z. Hu. Multi-step-ahead time series prediction using multiple-output support vector regression. *Neurocomputing*, 129:482–493, 2014.

[17] Y. Bao, X. Zhao, L. Wang, W. Qian, and J. Sun. Morphology-based classification of mycobacteria-infected macrophages with convolutional neural network: reveal EsxA-induced morphologic changes indistinguishable by naked eyes. *Translational Research*, 212:1–13, 2019.

[18] L. Barghout and L. Lee. Perceptual information processing system. In *Google Patents – US 2004/0059754 A1*, pages 1–20. 2004.

[19] D. A. Barr, C. Schutz, A. Balfour, M. Shey, M. Kamariza, C. R. Bertozzi, T. J. de Wet, R. Dinkele, A. Ward, and K. A. Haigh. Serial measurement of M. tuberculosis in blood from critically-ill patients with HIV-associated tuberculosis. *EBioMedicine*, 78(1):103949, 2022.

[20] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[21] A. Beykikhoshk, O. Arandjelović, D. Phung, and S. Venkatesh. Overcoming data scarcity of Twitter: using tweets as bootstrap with application to autism-related topic content analysis. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 1354–1361, 2015.

[22] C. M. Bishop and N. M. Nasrabadi. Regularization in Neural Networks. In *Pattern recognition and machine learning*, volume 4, pages 259–260. 2006.

[23] L. Boeck, S. Burbaud, M. Skwark, W. H. Pearson, J. Sangen, A. W. Wuest, E. K. P. Marshall, A. Weimann, I. Everall, J. M. Bryant, S. Malhotra, B. P. Bannerman, K. Kierdorf, T. L. Blundell, M. S. Dionne, J. Parkhill, and R. Andres Floto. Mycobacterium abscessus pathogenesis identified by phenogenomic analyses. *Nature Reviews Microbiology*, 7(9):1431–1441, 2022.

[24] C. C. Boehme, P. Nabeta, D. Hillemann, M. P. Nicol, S. Shenai, F. Krapp, J. Allen, R. Tahirli, R. Blakemore, R. Rustomjee, A. Milovic, M. Jones, S. M. O'Brien, D. H. Persing, S. Ruesch-Gerdes, E. Gotuzzo, C. Rodrigues, D. Alland, and M. D. Perkins. Rapid molecular detection of tuberculosis and rifampin resistance. *New England Journal of Medicine*, 363(11):1005–1015, 2010.

[25] Marc-O. Boldi, J. Denis-Lessard, R. Neziri, R. Brouillet, C. Von-Garnier, V. Chavez, J. Mazza-Stalder, K. Jaton, G. Greub, and O. Opota. Performance of microbiological tests for tuberculosis diagnostic according to the type of respiratory specimen: A 10-year retrospective study. *Frontiers in Cellular and Infection Microbiology*, 13:1131241, 2023.

[26] L. J. Bonnett and G. R. Davies. Quality of outcome reporting in phase II studies in pulmonary tuberculosis. *Trials*, 16:518, 2015.

[27] L. J. Bonnett, G. Ken-Dror, and G. R. Davies. Quality of reporting of outcomes in phase III studies of pulmonary tuberculosis: a systematic review. *Trials*, 19(1):134, 2018.

[28] L. J. Bonnett, G. Ken-Dror, G. C. K. W. Koh, and G. R. Davies. Comparing the efficacy of drug regimens for pulmonary tuberculosis: meta-analysis of endpoints in early-phase clinical trials. *Clinical Infectious Diseases*, 65(1):46–54, 2017.

[29] H. Bouma, A. Vilanova, J. O. Bescós, B. M ter Haar Romeny, and F. A. Gerritsen. Fast and accurate Gaussian derivatives based on B-splines. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 406–417, 2007.

[30] B. Braden. The surveyor's area formula. *The College Mathematics Journal*, 17(4):326–337, 1986.

[31] John Canny. A Computational Approach to Edge Detection. *Transactions on Pattern Analysis and Machine Intelligence*, 6:679–698, 1986.

[32] S. Chakravorty, A. M. Simmons, M. Rowneki, H. Parmar, Y. Cao, J. Ryan, P. P. Banada, S. Deshpande, S. Shenai, A. Gall, J. Glass, B. Krieswirth, S. G. Schumacher, P. Nabeta, N. Tukvadze, C. Rodrigues, A. Skrahina, E. Tagliani, D. M. Cirillo, A. Davidow, C. M. Denkinger, D. Persing, R. Kwiatkowski, M. Jones, and D. Alland. The New Xpert MTB/RIF Ultra: Improving Detection of Mycobacterium tuberculosis and Resistance to Rifampin in an Assay Suitable for Point-of-Care Testing. *mBio*, 8(4):e00812–17, 2017.

[33] V. Chalana and Y. Kim. A methodology for evaluation of boundary detection algorithms on medical images. *Transactions on Medical Imaging*, 16(5):642–652, 1997.

[34] J. Chang, P. Arbeláez, N. Switz, C. Reber, A. Tapley, J. L. Davis, A. Cattamanchi, D. Fletcher, and J. Malik. Automated tuberculosis diagnosis using fluorescence images from a mobile microscope. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 345–352, 2012.

[35] M. Chaplin, J. J. Kirkham, K. Dwan, D. J. Sloan, G. Davies, and A. L. Jorgensen. STrengthening the reporting of pharmacogenetic studies: Development of the STROPS guideline. *PLoS Medicine*, 17(9):e1003344, 2020.

[36] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *British Machine Vision Conference*, 2012. `doi:10.5244/C.26.21`.

[37] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *International Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[38] Yi. Chi, Z. Xiong, Q. Chang, C. Li, and H. Sheng. Improving Hessian matrix detector for SURF. *Transactions on Information and Systems*, 94(4):921–925, 2011.

[39] E. S. Chung, W. C. Johnson, and B. B. Aldridge. Types and functions of heterogeneity in mycobacteria. *Nature Reviews Microbiology*, 20(9):529–541, 2022.

[40] A. J. Codlin, T. P. Dao, L. N. Q. Vo, R. J. Forse, V. Van Truong, H. M. Dang, L. H. Nguyen, H. B. Nguyen, N. V. Nguyen, K. Sidney-Annerstedt, B. Squire, K. Lönnroth, and M. Caws. Independent evaluation of 12 artificial intelligence solutions for the detection of tuberculosis. *Scientific Reports*, 11(1):23895, 2021.

[41] L. E. Connolly, P. H. Edelstein, and L. Ramakrishnan. Why is long-term therapy required to cure tuberculosis? *PLoS Medicine*, 4(3):e120, 2007.

[42] F. Conradie, T. R. Bagdasaryan, S. Borisov, P. Howell, L. Mikiashvili, N. Ngubane, A. Samoilova, S. Skornykova, E. Tudor, E. Variava, P. Yablonskiy, D. Everitt, G. H. Wills, E. Sun, M. Olugbosi, E. Egizi, M. Li, A. Holsta, J. Timm, A. Bateson, A. M. Crook, S. M. Fabiane, R. Hunt, T. D. McHugh, C. D. Tweed, S. Foraida, C. M. Mendel, and M. Spigelman. Bedaquiline–Pretomanid–Linezolid regimens for drug-resistant tuberculosis. *New England Journal of Medicine*, 387(9):810–823, 2022.

[43] F. Conradie, A. H. Diacon, N. Ngubane, P. Howell, D. Everitt, A. M. Crook, C. M. Mendel, E. Egizi, J. Moreira, J. Timm, T. D. McHugh, G. H. Wills, A. Bateson, R. Hunt, C. Van Niekerk, M. Li, M. Olugbosi, and M. Spigelman. Treatment of highly drug-resistant pulmonary tuberculosis. *New England Journal of Medicine*, 382(10):893–902, 2020.

[44] J. Cooper, O. Arandjelović, and D. J. Harrison. Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping. *Pattern Recognition*, 129:108743, 2022.

[45] J. Cooper, I. H. Um, O. Arandjelović, and D. J. Harrison. Lymphocyte Classification from Hoechst Stained Slides with Deep Learning. *Cancers*, 14(23):5957, 2021.

[46] E. L. Corbett, C. J. Watt, N. Walker, D. Maher, B. G. Williams, M. C. Raviglione, and C. Dye. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Archives of Internal Medicine*, 163(9):1009–1021, 2003.

[47] M. G. F. Costa, C. F. F. Costa Filho, A. Kimura, P. C. Levy, C. M. Xavier, and L. B. Fujimoto. A sputum smear microscopy image database for automatic bacilli detection in conventional microscopy. In *Annual International Conference of the Engineering in Medicine and Biology Society*, pages 2841–2844, 2014.

[48] M. G. F. Costa, C. F. F. Costa Filho, J. F. Sena, J. Salem, and M. O. de Lima. Automatic identification of mycobacterium tuberculosis with conventional light microscopy. In *International Conference of Engineering in Medicine and Biology Society*, pages 382–385, 2008.

[49] C. F. F. Costa Filho, M. G. F. Costa, and A. K. Júnior. Autofocus functions for tuberculosis diagnosis with conventional sputum smear microscopy. *Current Microscopy Contributions to Advances in Science and Technology*, 31(1):13–20, 2012.

[50] C. F. F. Costa Filho, P. C. Levy, C. de M. Xavier, L. B. M. Fujimoto, and M. G. F. Costa. Automatic identification of tuberculosis mycobacterium. *Research on Biomedical Engineering*, 31:33–43, 2015.

[51] C. F. F. CostaFilho, P. C. Levy, C. M. Xavier, M. G. F. Costa, L. B. M. Fujimoto, and J. Salem. Mycobacterium tuberculosis recognition with conventional microscopy. In *International Conference of the Engineering in Medicine and Biology Society*, pages 6263–6268, 2012.

[52] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[53] Jaiyanth Daniel, Hédia Maamar, Chirajyoti Deb, Tatiana D. Sirakova, and Pappachan E. Kolattukudy. Mycobacterium tuberculosis uses host triacylglycerol to accumulate lipid droplets and acquires a dormancy-like phenotype in lipid-loaded macrophages. *PLoS Pathogens*, 7(6):e1002093, 2011.

[54] T. M. Daniel. The history of tuberculosis. *Respiratory Medicine*, 100(11):13–24, 2006.

[55] K. Daniel Chaves Viquez, O. Arandjelovic, A. Blaikie, and I. Ae Hwang. Synthesising wider field images from narrow-field retinal video acquired using a low-cost direct ophthalmoscope (Arclight) attached to a smartphone. In *International Conference on Computer Vision Workshops*, pages 90–98, 2017.

[56] B. M. de Vries, G. J. C. Zwezerijnen, G. L. Burchell, F. H. P. van Velden, and R. Boellaard. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: a literature review. *Frontiers in Medicine*, 10:1180773, 2023.

[57] C. Deb, C. M. Lee, V. S. Dubey, J. Daniel, B. Abomoelak, T. D. Sirakova, S. Pawar, L. Rogers, and P. E. Kolattukudy. A Novel In Vitro Multiple-Stress Dormancy Model for Mycobacterium tuberculosis Generates a Lipid-Loaded, Drug-Tolerant, Dormant Pathogen. *PLoS One*, 4(6):e6077, 2009.

[58] S. Devadatta, P. R. Gangadharam, R. H. Andrews, W. Fox, C. V. Ramakrishnan, J. B. Selkon, and S. Velu. Peripheral neuritis due to isoniazid. *Bull World Health Organ*, 23(4-5):587–598, 1960.

[59] K. Dheda, T. Perumal, H. Moultrie, R. Perumal, A. Esmail, A. J. Scott, Z. Udwadia, K. C. Chang, J. Peter, A. Pooran, A. von Delft, D. von Delft, N. Martinson, M. Loveday, S. Charalambous, E. Kachingwe, W. Jassat, C. Cohen, S. Tempia, K. Fennelly, and M. Pai. The intersecting pandemics of tuberculosis and COVID-19: population-level and patient-level impact, clinical presentation, and corrective interventions. *The Lancet Respiratory Medicine*, 10(6):603–622, 2022.

[60] J. L. Díaz-Huerta, A. del Carmen Téllez-Anguiano, M. Fraga-Aguilar, J. A. Gutierrez-Gnecchi, and S. Arellano-Calderón. Image processing for AFB segmentation in bacilloscopies of pulmonary tuberculosis diagnosis. *PLoS One*, 14(7):e0218861, 2019.

[61] R. P. Dickson and G. B. Huffnagle. The lung microbiome: new principles for respiratory bacteriology in health and disease. *PLoS Pathogens*, 11(7):e1004923, 2015.

[62] N. Dimitriou, O. Arandjelović, and P. D. Caie. Deep learning for whole slide image analysis: an overview. *Frontiers in Medicine*, 6:264, 2019.

[63] R. Dinesh Jackson Samuel and B. Rajesh Kanna. Tuberculosis detection system using deep neural networks. *Neural Computing and Applications*, 31(5):1533–1545, 2018.

[64] H. D. Donoghue, M. Spigelman, C. L. Greenblatt, G. Lev-Maor, G. Kahila Bar-Gal, C. Matheson, K. Vernon, A. G Nerlich, and A. R Zink. Tuberculosis: from prehistory to Robert Koch, as revealed by ancient DNA. *The Lancet Infectious Diseases*, 4(9):584–592, 2004.

[65] S. E. Dorman, P. Nahid, E. V. Kurbatova, Patrick P. J. Phillips, K. Bryant, K. E. Dooley, M. Engle, S. V. Goldberg, H. T. T. Phan, J. Hakim, J. L. Johnson, M. Lourens, N. A. Martinson, G. Muzanyi, K. Narunsky, S. Nerette, N. V. Nguyen, T. H. Pham, S. Pierre, A. E. Purfield, W. Samaneka, R. M. Savic, I. Sanne, N. A. Scott, J. Shenje, E. Sizemore, A. Vernon, Z. Waja, M. Weiner, S. Swindells, and R. E. Chaisson. Four-month rifapentine regimens with or without moxifloxacin for tuberculosis. *New England Journal of Medicine*, 384(18):1705–1718, 2021.

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, pages 1–22, 2021.

[67] D. H. Douglas and T. K. Peucker. 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, 1973.

[68] T. J. S. Durant, S. N. Dudgeon, J. McPadden, A. Simpson, N. Price, W. L. Schulz, R. Torres, and E. M. Olson. Applications of digital microscopy and densely connected convolutional neural networks for automated quantification of babesia-infected erythrocytes. *Clinical Chemistry*, 68(1):218–229, 2022.

[69] C. Dye and B. G. Williams. The population dynamics and control of tuberculosis. *Science*, 328(5980):856–861, 2010.

[70] O. P. Dzyubak and E. L. Ritman. Automation of hessian-based tubularity measure response function in 3D biomedical images. *International Journal of Biomedical Imaging*, 2011:920401, 2011.

[71] M. El-Melegy, D. Mohamed, T. ElMelegy, and M. Abdelrahman. Identification of tuberculosis bacilli in ZN-stained sputum smear images: a deep learning approach. In *Conference on Computer Vision and Pattern Recognition Workshops*, pages 1131–1137, 2019.

[72] J. Fan and O. Arandjelović. Employing domain specific discriminative information to address inherent limitations of the LBP descriptor in face recognition. In *International Joint Conference on Neural Networks*, pages 1–7, 2018.

[73] M. Forero, G. Cristobal, and J. Alvarez-Borrego. Automatic identification techniques of tuberculosis bacteria. In *Applications of Digital Image Processing*, pages 71–81, 2003.

[74] M. G. Forero, G. Cristóbal, and M. Desco. Automatic identification of Mycobacterium tuberculosis by Gaussian mixture models. *Journal of Microscopy*, 223(2):120–132, 2006.

[75] M. G. Forero, F. Sroubek, and G. Cristóbal. Identification of tuberculosis bacteria based on shape and color. *Real-Time Imaging*, 10(4):251–262, 2004.

[76] M.G Forero-Vargas, F. Sroubek, J. Alvarez-Borrego, N. Malpica, G. Cristóbal, A. Santos, L. Alcalá, M. Desco, and L. Cohen. Segmentation, autofocusing, and signature extraction of tuberculosis sputum images. In *Photonic Devices and Algorithms for Computing IV*, pages 171–182, 2002.

[77] T. T. Frie, N. Cristianini, and C. Campbell. The kernel-adatron algorithm: a fast and simple learning procedure for support vector machines. In *Machine Learning*, pages 188–196, 1998.

[78] S. O. Friedrich, A. Rachow, E. Saathoff, K. Singh, C. D. Mangu, R. Dawson, P. P. Phillips, A. Venter, A. Bateson, C. C. Boehme, N. Heinrich, R. D. Hunt, M. J. Boeree, A. Zumla, T. D. McHugh, S. H. Gillespie, A. H. Diacon, and M. Hoelscher. Assessment of the sensitivity and specificity of Xpert MTB/RIF assay as an early sputum biomarker of response to tuberculosis treatment. *The Lancet Respiratory Medicine*, 1(6):462–470, 2013.

[79] M. Gadermayr, L. Gupta, B. M. Klinkhammer, P. Boor, and D. Merhof. Unsupervisedly training GANs for segmenting digital pathology with automatically generated annotations. In *Machine Learning Research*, pages 175–184, 2019.

[80] N. J. Garton, S. J. Waddell, A. L. Sherratt, Su-M. Lee, R. J. Smith, C. Senner, J. Hinds, K. Rajakumar, R. A. Adegbola, G. S. Besra, P. D. Butcher, and M. R. Barer. Cytological and Transcript Analyses Reveal Fat and Lazy Persister-Like Bacilli in Tuberculous Sputum. *PLoS Medicine*, 5(4):e75, 2008.

[81] A. A. Gele, G. Bjune, and F. Abebe. Pastoralism and delay in diagnosis of TB in Ethiopia. *BMC Public Health*, 9(1):1009–1021, 2009.

[82] R. S. Ghiass, O. Arandjelovic, H. Bendada, and X. Maldague. Vesselness features and the inverse compositional AAM for robust face recognition using thermal IR. In *Conference on Artificial Intelligence*, pages 357–364, 2013.

[83] P. Ghosh, D. Bhattacharjee, and M. Nasipuri. A hybrid approach to diagnosis of tuberculosis from sputum. In *International Conference on Electrical, Electronics, and Optimization Techniques*, pages 771–776, 2016.

[84] D. J. Girling. The hepatic toxicity of antituberculosis regimens containing isoniazid, rifampicin and pyrazinamide. *Tubercle*, 59(1):13–32, 1977.

[85] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, pages 580–587, 2014.

[86] A. Glassing, S. E. Dowd, S. Galandiuk, B. Davis, and R. J. Chiodini. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathogens*, 8:1–12, 2016.

[87] R. C. Gonzalez, R. E. Woods, and B. R. Masters. Digital image processing. *Journal of Biomedical Optics*, 14(2):029901, 2009.

[88] I. Goodfellow, Y. Bengio, and A. Courville. Chapter 5 on Practical Methodology. In *Deep Learning*, pages 120–123. 2016.

[89] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. In *Advances in Neural Information Processing Systems*, pages 139–144, 2014.

[90] P. Goodwin and R. Lawton. On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, 15(4):405–408, 1999.

[91] L. Govindan, N. Padmasini, and M. Yacin. Automated tuberculosis screening using Zeihl Neelson image. In *International Conference on Engineering and Technology*, pages 1–4, 2015.

[92] B. F. Green. Public Health Image Library (PHIL). *Bulletin of the Medical Library Association*, 89(2):243, 2001.

[93] B. Guindon and Y. Zhang. Application of the dice coefficient to accuracy assessment of object-based image classification. *Canadian Journal of Remote Sensing*, 43(1):48–61, 2017.

[94] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[95] M. Hailemariam and E. Azerefegne. Evaluation of Laboratory Professionals on AFB Smear Reading at Hawassa District Health Institutions, Southern Ethiopia. *International Journal of Research Studies in Microbiology and Biotechnology*, 4(4):12–19, 2018.

[96] R. J. H. Hammond, V. O. Baron, K. Oravcova, S. Lipworth, and S. H. Gillespie. Phenotypic resistance in mycobacteria: is it because I am old or fat that I resist you? *Journal of Antimicrobial Chemotherapy*, 70(10):2823–2827, 2015.

[97] R. J. H. Hammond, F. Kloprogge, O D. Pasqua, and S. H. Gillespie. Implications of drug-induced phenotypical resistance: Is isoniazid radicalizing M. tuberculosis? *Frontiers in Antibiotics*, 1:928365, 2022.

[98] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T. Y. Liu, and W. Y. Ma. Dual learning for machine translation. In *International Conference on Neural Information Processing Systems*, volume 29, pages 820–828, 2016.

[99] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[100] R D Hjelm, A P Jacob, T Che, A Trischler, K Cho, and Y Bengio. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*, 2017.

[101] T. C. Hollon, B. Pandian, A. R. Adapa, E. Urias, A V. Save, Siri Sahib S Khalsa, D. G. Eichberg, R. S. D'Amico, Z. U. Farooq, S. Lewis, P. D. Petridis, T. Marie, A. H. Shah, H. J. L. Garton, C. O. Maher, J. A. Heth, E. L. McKean, S. E. Sullivan, S. L. Hervey-Jumper, P. G. Patil, B. G. Thompson, O. Sagher, G. M. McKhann, R. J. Komotar, M. E. Ivan, M. Snuderl, M. L. Otten, T. D. Johnson, M. B. Sisti, J. N. Bruce, K. M. Muraszko, J. Trautman, C. W. Freudiger, P. Canoll, H. Lee, S. Camelo-Piragua, and D. A. Orringer. Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nature Medicine*, 26(1):52–58, 2020.

[102] R. M. G. J. Houben and P. J. Dodd. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLoS Med*, 13(10):e1002152, 2016.

[103] M. Hu, Y. Liu, Y. Zhang, T. Guan, and Y. He. Automatic detection of tuberculosis bacilli in sputum smear scans based on subgraph classification. In *International Conference on Medical Imaging Physics and Engineering*, pages 1–7, 2019.

[104] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely Connected Convolutional Networks. In *Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[105] Z. Huang and J. Leng. Analysis of Hu's moment invariants on image scaling and rotation. In *International Conference on Computer Engineering and Technology*, page 476, 2010.

[106] R. C. Huard, M. Fabre, P. de Haas, L. Claudio Oliveira Lazzarini, D. van Soolingen, D. Cousins, and J. L. Ho. Novel genetic polymorphisms that further delineate the phylogeny of the mycobacterium tuberculosis complex. *Journal of Bacteriology*, 188(12):4271–4287, 2006.

[107] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*, 2016.

[108] M. Z. Imperial, P. Nahid, P. P. J. Phillips, G. R. Davies, K. Fielding, D. Hanna, D. Hermann, R. S. Wallis, J. L. Johnson, and C. Lienhardt. A patient-level pooled analysis of treatment-shortening regimens for drug-susceptible pulmonary tuberculosis. *Nature Medicine*, 24(11):1708–1715, 2018.

[109] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[110] G. James, D. Witten, T. Hastie, and R. Tibshirani. 7.8 Lab: non-linear modeling. In *An introduction to statistical learning*, pages 287–297. 2013.

[111] P. J. Jenner, G. A. Ellard, W. G. Allan, D. Singh, D. J. Girling, and A. J. Nunn. Serum uric acid concentrations and arthralgia among patients treated with pyrazinamide-containing regimens in Hong Kong and Singapore. *Tubercle*, 62(3):175–179, 1981.

[112] Canny John. A Computational Approach to Edge Detection. *Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8:679–698, 1986.

[113] M. I. Jordan and T. M. Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[114] S. Kant and M. M. Srivastava. Towards Automated Tuberculosis detection using Deep Learning. In *Symposium Series on Computational Intelligence*, pages 1250–1253, 2019.

[115] R. Khutlang, S. Krishnan, R. Dendere, A. Whitelaw, K. Veropoulos, G. Learmonth, and T. S. Douglas. Classification of Mycobacterium tuberculosis in images of ZN-stained sputum smears. *Transactions on Information Technology in Biomedicine*, 14(4):949–957, 2010.

[116] R. Khutlang, S. Krishnan, A. Whitelaw, and T. S. Douglas. Automated detection of tuberculosis in Ziehl-Neelsen-stained sputum smears using two one-class classifiers. *Journal of Microscopy*, 237(1):96–102, 2010.

[117] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[118] B. Kouriba, O. Ouwe Missi Oukem-Boyer, B. Traoré, A. Touré, L. Raskine, and F. X. Babin. Installing biosafety level 3 containment laboratories in low- and middle-income countries: challenges and prospects from Mali's experience. *New Microbes New Infect*, 26:S74–S77, 2018.

[119] N. C. S. Kumar and Y. Radhika. Optimized maximum principal curvatures based segmentation of blood vessels from retinal images. *Biomedical Research*, 30(2):308–318, 2019.

[120] T. K Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, 1998.

[121] S. D. Lawn and A. I. Zumla. Tuberculosis. *The Lancet*, 378(9785):57–72, 2011.

[122] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[123] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, Al. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017.

[124] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, pages 1324–1332, 2010.

[125] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716, 2016.

[126] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li. Dice loss for data-imbalanced NLP tasks. In *Annual Meeting of the Association for Computational Linguistics*, pages 465–476, 2020.

[127] S. Liang, J. Ma, G. Wang, J. Shao, J. Li, H. Deng, C. Wang, and W. Li. The application of artificial intelligence in the diagnosis and drug resistance prediction of pulmonary tuberculosis. *Frontiers in Medicine*, 9:935080, 2022.

[128] P. L. Lin and J. L. Flynn. The end of the binary era: revisiting the spectrum of tuberculosis. *The Journal of Immunology*, 201(9):2541–2548, 2018.

[129] M. Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 469–477, 2016.

[130] Y. Liu, G. Cao, Z. Ge, and Y. Hu. Crowd counting method via a dynamic-refined density map network. *Neurocomputing*, 497:191–203, 2022.

[131] M. M. Logsdon and B. B. Aldridge. Stable Regulation of Cell Cycle Events in Mycobacteria: Insights From Inherently Heterogeneous Bacterial Populations. *Frontiers in Microbiology*, 9:514, 2018.

[132] A. Lomacenkova and O. Arandjelović. Whole slide pathology image patch based deep classification: an investigation of the effects of the latent autoencoder representation and the loss function form. In *International Conference on Biomedical and Health Informatics*, pages 1–4, 2021.

[133] K. Lönnroth, L. M. Thuong, P. D. Linh, and V. K. Diwan. Delay and discontinuity – A survey of TB patients' search of a diagnosis in a diversified health care system. *International Journal of Tuberculosis and Lung Disease*, 3(11):992–1000, 1999.

[134] Y.P. López, C. F. F. Costa F., L. M. R. Aguilera, and M. G. F. Costa. Automatic classification of light field smear microscopy patches using convolutional neural networks for identifying mycobacterium tuberculosis. In *Chilean Conference on Electrical, Electronics Engineering, Information and Communication Technologies*, pages 1895–1898, 2017.

[135] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference of Learning Representations*, pages 1–16, 2016.

[136] D. G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, 1999.

[137] V. Makkapati, R. Agrawal, and R. Acharya. Segmentation and classification of tuberculosis bacilli from ZN-stained sputum smear images. In *International Conference on Automation Science and Engineering*, pages 217–220, 2009.

[138] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *EUROSPEECH*, pages 1–4, 1997.

[139] American Journal of Respiratory Medicine and Critical Care. Diagnostic Standards and Classification of Tuberculosis in Adults and Children. *American Journal of Respiratory and Critical Care Medicine*, 161(4):1376–1395, 2000.

[140] P. K. Mehta, A. Raj, N. Singh, and G. K. Khuller. Diagnosis of extrapulmonary tuberculosis by PCR. *FEMS Immunology & Medical Microbiology*, 66(1):20–36, 2012.

[141] F. A. Merchant and K. R. Castleman. Computer-assisted microscopy. In *The Essential Guide to Image Processing*, pages 777–831. 2009.

[142] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey. *Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.

[143] M. Mirza and S Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

[144] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. In *International Conference on Neural Information Processing Systems*, pages 4829–4840, 2017.

[145] D. A. Mitchison. Basic mechanisms of chemotherapy. *Chest*, 76(6 Suppl):771–781, 1979.

[146] D. A. Mitchison. The search for new sterilizing anti-tuberculosis drugs. *Frontiers in Bioscience*, 9(1-3):1059, 2004.

[147] K. S. Mithra and W. R. Sam Emmanuel. FHDT: fuzzy and Hyco-entropy-based decision tree classifier for tuberculosis diagnosis from sputum images. *Sādhanā*, 43(8):125, 2018.

[148] K. S. Mithra and W. R. Sam Emmanuel. Automated identification of mycobacterium bacillus from sputum images for tuberculosis diagnosis. *Signal, Image and Video Processing*, 13(8):1585–1592, 2019.

[149] B. Mtafya, I. Sabi, J. John, E. Sichone, W. Olomi, S. H. Gillespie, N. E. Ntinginya, and W. Sabiiti. Systematic assessment of clinical and bacteriological markers for tuberculosis reveals discordance and inaccuracy of symptom-based diagnosis for treatment response monitoring. *Frontiers in Medicine*, 9(28):3096, 2022.

[150] A. Mukundan, G. Tolias, A. Bursuc, H. Jégou, and O. Chum. Understanding and improving kernel local descriptors. *International Journal of Computer Vision*, 127(11):1723–1737, 2019.

[151] C. J. F. Mundy, I. Bates, W. Nkhomal, K. Floyd, G. Kadewele, M. Ngwira, A. Khuwi, S. B. Squire, and C. F. Gilks. The operation, quality and costs of a district hospital laboratory service in Malawi. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 97(4):403–408, 2003.

[152] E. J. Muñoz-Elías and J. D. McKinney. Mycobacterium tuberculosis isocitrate lyases 1 and 2 are jointly required for in vivo growth and virulence. *Nature Medicine*, 11(6):638–644, 2005.

[153] R. Nayak, V. P. Shenoy, and R. R. Galigekere. A new algorithm for automatic assessment of the degree of TB-infection using images of ZN-stained sputum smear. In *International Conference on Systems in Medicine and Biology*, pages 294–299, 2010.

[154] A. Newell. A Tutorial on Speech Understanding Systems. *Speech Recognition*, pages 4–54, 1975.

[155] M. Nicas, W. W. Nazaroff, and A. Hubbard. Toward understanding the risk of secondary airborne infection: emission of respirable pathogens. *Journal of Occupational and Environmental Hygiene*, 2(3):143–154, 2005.

[156] M. Niemi, J. T. Backman, M. F. Fromm, P. J. Neuvonen, and K. T. Kivisttö. Pharmacokinetic Interactions with Rifampicin. *Clinical Pharmacokinetics*, 42(9):819–850, 2003.

[157] Bern-T. Nyang'wa, C. Berry, E. Kazounis, I. Motta, N. Parpieva, Z. Tigay, V. Solodovnikova, I. Liverko, R. Moodliar, M. Dodd, N. Ngubane, M. Rassool, T. D. McHugh, M. Spigelman,

D. A. J. Moore, K. Ritmeijer, P. du Cros, and K. Fielding. A 24-week, all-oral regimen for rifampin-resistant tuberculosis. *New England Journal of Medicine*, 387(25):2331–2343, 2022.

[158] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *International Conference on Pattern Recognition*, pages 582–585, 1994.

[159] N. O'Mahony, S. Campbell, A. Carvalho, S. Harapanahalli, G. V. Hernandez, L. Krpalkova, D. Riordan, and J. Walsh. Deep learning vs. traditional computer vision. In *Science and Information Conference*, pages 128–144, 2019.

[160] World Health Organization. TB: A Global Emergency – WHO Report on the TB Epidemic, 1994.

[161] World Health Organization. Global Tuberculosis Report 2006. Technical report, World Health Organization, Geneva, 2006.

[162] World Health Organization. End TB Strategy, 2015. URL: `https://apps.who.int/iris/bitstream/handle/10665/331326/WHO-HTM-TB-2015.19-eng.pdf?sequence=1&isAllowed=y`.

[163] World Health Organization. Global Tuberculosis Report 2020. Technical report, World Health Organization, Geneva, 2020.

[164] World Health Organization. Global Tuberculosis Report 2021. Technical report, World Health Organization, Geneva, 2021.

[165] World Health Organization. Global Tuberculosis Report 2022. Technical report, World Health Organization, Geneva, 2022.

[166] World Health Organization. WHO consolidated guidelines on tuberculosis. Module 4: Treatment. Drug-resistant tuberculosis treatment, 2022. URL: `https://www.who.int/publications/i/item/9789240063129`.

[167] World Health Organization. WHO operational handbook on tuberculosis. Module 4: Treatment. Drug-susceptible tuberculosis treatment, 2022. URL: `https://www.who.int/publications/i/item/9789240050761`.

[168] M. K. O'Shea, G. C. K. W. Koh, M. Munang, G. Smith, A. Banerjee, and M. Dedicoat. Time-to-detection in culture predicts risk of mycobacterium tuberculosis transmission: a cohort study. *Clinical Infectious Diseases*, 59(2):177–185, 2014.

[169] M. K. Osman, F. Ahmad, Z. Saad, M. Y. Mashor, and H. Jaafar. A genetic algorithm-neural network approach for mycobacterium tuberculosis detection in Ziehl-Neelsen stained tissue slide images. In *International Conference on Intelligent Systems Design and Applications*, pages 1229–1234, 2010.

[170] M. K. Osman, M. Y. Mashor, and H. Jaafar. Detection of mycobacterium tuberculosis in Ziehl-Neelsen stained tissue images using Zernike moments and hybrid multilayered perceptron network. In *International Conference on Systems, Man and Cybernetics*, pages 4049–4055, 2010.

[171] M. K. Osman, M. Y. Mashor, Z Saad, and H Jaafar. Colour image segmentation of tuberculosis bacilli in Ziehl-Neelsen-stained tissue images using moving k-mean clustering procedure. In *Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation*, pages 215–220, 2010.

[172] M Pai, M A Behr, D Dowdy, K Dheda, M Divangahi, C C Boehme, A Ginsberg, S Swaminathan, M Spigelman, H Getahun, D Menzies, and M Raviglione. Tuberculosis. *Nature Reviews Disease Primers*, 2:16076, 2016.

[173] J C Palomino. Nonconventional and new methods in the diagnosis of tuberculosis: feasibility and applicability in the field. *European Respiratory Journal*, 26(2):339–350, 2005.

[174] R. O. Panicker, K. S. Kalmady, J. Rajan, and M. K. Sabu. Automatic detection of tuberculosis bacilli from microscopic sputum smear images using deep learning methods. *Biocybernetics and Biomedical Engineering*, 38(3):691–699, 2018.

[175] Stop T B Partnership. Global Laboratory Initiative advancing TB diagnosis: mycobacteriology laboratory manual, 2014. URL: `https://stoptb.org/wg/gli/assets/documents/gli_mycobacteriology_lab_manual_web.pdf`.

[176] Y. Payasi and S. Patidar. Diagnosis and counting of tuberculosis bacilli using digital image processing. In *International Conference on Information, Communication, Instrumentation and Control*, pages 1–5, 2017.

[177] A. Penn-Nicholson, S. B. Georghiou, N. Ciobanu, M. Kazi, M. Bhalla, A. David, F. Conradie, M. Ruhwald, V. Crudu, C. Rodrigues, V. P. Myneedu, L. Scott, C. M. Denkinger, and S. G. Schumacher. Detection of isoniazid, fluoroquinolone, ethionamide, amikacin, kanamycin, and capreomycin resistance by the Xpert MTB/XDR assay: a cross-sectional multicentre diagnostic accuracy study. *The Lancet Infectious Diseases*, 22(2):242–249, 2022.

[178] W. K. Pratt. *Digital image processing: PIKS Scientific Inside*, volume 4. 2007.

[179] E. Priya and S. Srinivasan. Automated object and image level classification of TB images using support vector neural network classifier. *Biocybernetics and Biomedical Engineering*, 36(4):670–678, 2016.

[180] M. Pultar. Improving the HardNet Descriptor. *arXiv preprint arXiv:2007.09699*, 2020.

[181] Y. Qian, L. Zhang, X. Hong, C. Donovan, and O. Arandjelovic. Segmentation assisted u-shaped multi-scale transformer for crowd counting. In *British Machine Vision Conference*, page 397, 2022.

[182] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation learning with Deep Convolutional GANs. *arXiv preprint arXiv:1511.06434*, 2015.

[183] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2017.

[184] K. Richardson, O. T. Bennion, S. Tan, A. N. Hoang, M. Cokol, and B. B. Aldridge. Temporal and intrinsic factors of rifampicin tolerance in mycobacteria. *Proceedings of the National Academy of Sciences*, 113(29):8302–8307, 2016.

[185] H. L. Rieder, A. Van Deun, K. Man Kam, S. Jae Kim, T. M. Chonde, A. Trebucq, and R. Urbanczik. *Priorities for Tuberculosis Bacteriology Services in Low-Income Countries*. International Union Against Tuberculosis and Lung Disease, 2007.

[186] M. A. Riojas, K. J. McGough, C. J. Rider-Riojas, N. Rastogi, and M. H. Hazbón. Phylogenomic analysis of the species of the mycobacterium tuberculosis complex demonstrates that mycobacterium africanum, mycobacterium bovis, mycobacterium caprae, mycobacterium microti and mycobacterium pinnipedii are later heterotypic synon. *International Journal of Systematic and Evolutionary Microbiology*, 68(1):324–332, 2018.

[187] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[188] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[189] M. Rudzki. Vessel detection method based on eigenvalues of the hessian matrix and its applicability to airway tree segmentation. In *International PhD Workshop*, pages 100–105, 2009.

[190] W. Sabiiti, K. Azam, E. C. W. Farmer, D. Kuchaka, B. Mtafya, R. Bowness, K. Oravcova, I. Honeyborne, D. Evangelopoulos, T. D. McHugh, C. Khosa, A. Rachow, N. Heinrich, E. Kampira, G. Davies, N. Bhatt, E. N. Ntinginya, S. Viegas, I. Jani, M. Kamdolozi, A. Mdolo, M. Khonga, M. J. Boeree, P. P. J. Phillips, D. Sloan, M. Hoelscher, G. Kibiki, and S. H. Gillespie. Tuberculosis bacillary load, an early marker of disease severity: the utility of tuberculosis Molecular Bacterial Load Assay. *Thorax*, 75(7):606–608, 2020.

[191] K. S. Sachdeva and N. Kumar. Closing the gaps in tuberculosis detection—considerations for policy makers. *The Lancet Global Health*, 11(2):e185–e186, 2023.

[192] P. Sadaphal, J. Rao, G. W. Comstock, and M. F. Beg. Image processing techniques for identifying Mycobacterium tuberculosis in Ziehl-Neelsen stains. *The International Journal of Tuberculosis and Lung Disease*, 12(5):579–582, 2008.

[193] A. Sagar. ViTBIS: vision transformer for biomedical image segmentation. In *Clinical Image-Based Procedures, Distributed and Collaborative Learning, Artificial Intelligence for Combating COVID-19 and Secure and Privacy-Preserving Machine Learning*, pages 34–45, 2021.

[194] B. Said, L. Charlie, E. Getachew, C. L. Wanjiru, M. Abebe, and T. Manyazewal. Molecular bacterial load assay versus culture for monitoring treatment response in adults with tuberculosis. *SAGE Open Medicine*, 9:20503121211033470, 2021.

[195] H. Salem, D. Soria, J. N. Lund, and A. Awwad. A Systematic Review of the Applications of Expert Systems and Machine Learning in Clinical Urology. *BMC Medical Informatics and Decision Making*, 21:1–36, 2021.

[196] S. L. Salzberg. C4.5: Programs for Machine Learning. In *Machine Learning*, volume 16, pages 235–240. 1994.

[197] R. Santiago-Mozos, F. Pérez-Cruz, M. G. Madden, and A. Artés-Rodríguez. An automated screening system for tuberculosis. *Journal of Biomedical and Health Informatics*, 18(3):855–862, 2013.

[198] C. Sekaggya-Wiltshire, R. Nabisere, J. Musaazi, B. Otaalo, F. Aber, L. Alinaitwe, J. Nampala, L. Najjemba, A. Buzibye, D. Omali, K. Gausi, A. Kengo, M. Lamorde, R. Aarnoutse, P. Denti, K. E. Dooley, and D. J. Sloan. Decreased dolutegravir and efavirenz

concentrations with preserved virological suppression in patients with tuberculosis and human immunodeficiency virus receiving high-dose rifampicin. *Clinical Infectious Diseases*, 76(3):e910–e919, 2023.

[199] M. K. M. Serrão, M. G. F. Costa, L. B. Fujimoto, M. M. Ogusku, and C. F. F. Costa Filho. Automatic bacillus detection in light field microscopy images using convolutional neural networks and mosaic imaging approach. In *International Conference of Engineering in Medicine and Biology Society*, pages 1903–1906, 2020.

[200] M. I. Shah, S. Mishra, V. K. Yadav, A. Chauhan, M. Sarkar, S. K. Sharma, and C. Rout. Ziehl–Neelsen sputum smear microscopy image database: a resource to facilitate automated bacilli detection for tuberculosis diagnosis. *Journal of Medical Imaging*, 4(2):027503, 2017.

[201] L. Shapiro. *Computer vision and image processing*. Academic Press, 1992.

[202] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, pages 1–12, 2015.

[203] V. A. Sindagi and V. M. Patel. Generating high-quality crowd density maps using contextual pyramid CNNs. In *International Conference on Computer Vision*, pages 1861–1870, 2017.

[204] S. Singh. Emphasis on the Minimization of False Negatives or False Positives in Binary Classification. *arXiv preprint arXiv:2204.02526*, 2022.

[205] D. J. Sloan and M. Dedicoat. Tuberculosis management and drug resistance in Hlabisa hospital, KwaZulu Natal. *Journal of Infection*, 55(3):e39, 2007.

[206] D. J Sloan and J. M. Lewis. Management of multidrug-resistant TB: novel treatments and their expansion to low resource settings. *Transactions of The Royal Society of Tropical Medicine and Hygiene*, 110(3):163–172, 2016.

[207] D. J. Sloan, H. C. Mwandumba, N. J. Garton, S. H. Khoo, A. E. Butterworth, T. J. Allain, R. S. Heyderman, E. L. Corbett, M. R. Barer, and G. R. Davies. Pharmacodynamic modeling of bacillary elimination rates and detection of bacterial lipid bodies in sputum to predict and understand outcomes in treatment of pulmonary tuberculosis. *Clinical Infectious Diseases*, 61(1):1–8, 2015.

[208] L. N. Smith. Cyclical learning rates for training neural networks. In *Winter Conference on Applications of Computer Vision*, pages 464–472, 2017.

[209] R. S. Soans, V. P. Shenoy, and R. R. Galigekere. Automatic assessment of the degree of TB-infection using images of ZN-stained sputum smear: New results. In *International Conference on Systems in Medicine and Biology*, pages 22–25, 2016.

[210] I. Sobel and G. Feldman. A 3×3 isotropic gradient operator for image processing. In *Stanford Artificial Project*, pages 271–272, 1968.

[211] J. M. Sosa, D. E. Huber, B. Welk, and H. L. Fraser. Development and application of MIPAR™: a novel software package for two- and three-dimensional microstructural characterization. *Integrating Materials and Manufacturing Innovation*, 3(1):123–140, 2014.

[212] M. Sotaquira, L. Rueda, and R. Narvaez. Detection and quantification of bacilli and clusters present in sputum smear samples: a novel algorithm for pulmonary tuberculosis diagnosis. In *International Conference on Digital Image Processing*, pages 117–121. 2009.

[213] D. P.S. Spence, J. Hotchkiss, C. S.D. Williams, and P. D.O. Davies. Tuberculosis and poverty. *British Medical Journal*, 307(6907):759–761, 1993.

[214] K. R. Steingart, M. Henry, V. Ng, P. C. Hopewell, A. Ramsay, J. Cunningham, R. Urbanczik, M. Perkins, M. A. Aziz, and M. Pai. Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review. *Lancet Infectious Diseases*, 9(6):570–581, 2006.

[215] R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In *International Conference on Computer Vision*, pages 7262–7272, 2021.

[216] K Styblo and J Meijer. Impact of BCG vaccination programmes in children and young adults on the tuberculosis problem. *Tubercle*, 57(1):17–43, 1976.

[217] P. G. Suárez, C. J. Watt, E. Alarcón, J. Portocarrero, D. Zavala, R. Canales, F. Luelmo, M. A. Espinal, and C. Dye. The dynamics of tuberculosis in tesponse to 10 years of intensive control effort in peru. *The Journal of Infectious Diseases*, 184(4):473–8, 2001.

[218] I. Suleymanova, T. Balassa, S. Tripathi, C. Molnar, M. Saarma, Y. Sidorova, and P. Horvath. A deep convolutional neural network approach for astrocyte detection. *Scientific Reports*, 8(1):12878, 2018.

[219] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *International Conference on Neural Information Processing Systems*, pages 1988–1996, 2014.

[220] M. Surucu, Y. Isler, M. Perc, and R. Kara. Convolutional neural networks predict the onset of paroxysmal atrial fibrillation: Theory and applications. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(11):113119, 2021.

[221] K. Swetha, B. Sankaragomathi, and J. B. Thangamalar. Convolutional neural network based automated detection of mycobacterium bacillus from sputum images. In *International Conference on Inventive Computation Technologies*, pages 293–300, 2020.

[222] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[223] D. R. Tait, M. Hatherill, O. Van Der Meeren, A. M. Ginsberg, E. Van Brakel, B. Salaun, T. J. Scriba, E.J. Akite, H. M. Ayles, A. Bollaerts, M. A. Demoitié, A. Diacon, T. G. Evans, P. Gillard, E. Hellström, J. C. Innes, M. Lempicki, M. Malahleha, N. Martinson, D. Mesia Vela, M. Muyoyeta, V. Nduba, T. G. Pascal, M. Tameris, F. Thienemann, R. J. Wilkinson, and F. Roman. Final analysis of a trial of M72/AS01E vaccine to prevent tuberculosis. *New England Journal of Medicine*, 381(25):2429–2439, 2019.

[224] K. A. Talbert Estlin and A. A. Sadun. Risk factors for ethambutol optic toxicity. *International Ophthalmology*, 30(1):63–72, 2010.

[225] X. Tang. The role of artificial intelligence in medical imaging research. *BJR Open*, 2(1):20190031, 2019.

[226] N. R. Thanky, D. B. Young, and B. D. Robertson. Unusual features of the cell cycle in mycobacteria: polar-restricted growth and the snapping-model of cell division. *Tuberculosis*, 87(3):231–236, 2007.

[227] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Conference on Computer Vision and Pattern Recognition*, pages 11016–11025, 2019.

[228] J. T. Tippett, D. A. Borkowitz, L. C. Clapp, C. J. Koester, and A. Vanderburgh Jr. *Optical and electro-optical information processing*. Massachusetts Institute of Technology Cambridge, 1965.

[229] D. Tiwari and A. R. Martineau. Inflammation-mediated tissue damage in pulmonary tuberculosis and host-directed therapeutic strategies. *Seminars in Immunology*, 65:101672, 2023.

[230] K. Toman, T. R. Frieden, and World Health Organization. *Toman's tuberculosis: case detection, treatment and monitoring: questions and answers / edited by T Frieden*. World Health Organization, Geneva, 2nd edition, 2004.

[231] O. S. Toungoussova, G. Bjune, and D. A. Caugant. Epidemic of tuberculosis in the former soviet union: social and biological reasons. *Tuberculosis*, 86(1):1–10, 2006.

[232] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021.

[233] H. Trilaksana, G. N. Dwimudyari, A. S. Agoes, and D. B. Widhyatmoko. Sputum smear images database: A resource for deep learning study based to detect Bacilli for TB diagnose. In *International Conference on Physical Instrumentation and Advanced Materials*, pages 40013–40019, 2019.

[234] S. Uddin. Tuberculosis Image Dataset, 2018. URL: `https://www.kaggle.com/datasets/saife245/tuberculosis-image-datasets`.

[235] E. Ufimtseva, N. Eremeeva, D. Vakhrusheva, and S. Skornyakov. Mycobacterium tuberculosis shape and size variations in alveolar macrophages of tuberculosis patients. *European Respiratory Journal*, 54(63):PA4605, 2019.

[236] UN. United Nations Millennium Goals, 2000. URL: `http://www.un.org/millenniumgoals/aids.shtml`.

[237] S. Valsson and O. Arandjelović. Nuances of interpreting X-ray analysis by deep learning and lessons for reporting experimental findings. *Sci*, 4(1):3, 2022.

[238] S. G. van Dijk and M. M. Scheunemann. Deep learning for semantic segmentation on minimal hardware. In *Robot World Cup*, pages 349–361, 2018.

[239] S. Van Teeffelen, J. W. Shaevitz, and Z. Gitai. Image analysis in fluorescence microscopy: bacterial dynamics as a case study. *Bioessays*, 34(5):427–436, 2012.

[240] R. N. van Zyl-Smit, A. Binder, R. Meldau, H. Mishra, P.L. Semple, G. Theron, J. Peter, A. Whitelaw, S. K. Sharma, and R. Warren. Comparison of quantitative techniques including Xpert MTB/RIF to evaluate mycobacterial burden. *PloS One*, 6(12):e28815, 2011.

[241] D. Vente, O. Arandjelović, V. O. Baron, E. Dombay, and S. H. Gillespie. Using machine learning for automatic estimation of M. smegmatis cell count from fluorescence microscopy images. In *International Workshop on Health Intelligence*, pages 57–68, 2019.

[242] K. Veropoulos, G. Learmonth, C. Campbell, B. Knight, and J. Simpson. Automated identification of tubercle bacilli in sputum: A preliminary investigation. *Analytical and Quantitative Cytology and Histology*, 21(4):277–282, 1999.

[243] S. Vijay, D. N. Vinh, H. T. Hai, V. T. N. Ha, V. T. M. Dung, T. D. Dinh, H. N. Nhung, T. T. B. Tram, B. B. Aldridge, N. T. Hanh, D. D. A. Thu, N. H. Phu, G. E. Thwaites, and N. T. T. Thuong. Influence of Stress and Antibiotic Resistance on Cell-Length Distribution in Mycobacterium tuberculosis Clinical Isolates. *Frontiers in Microbiology*, 8:2296, 2017.

[244] E. Vynnycky and P. E. Fine. The annual risk of infection with mycobacterium tuberculosis in England and Wales since 1901. *The International Journal of Tuberculosis and Lung Disease*, 1(5):389–396, 1997.

[245] E. Walach and L. Wolf. Learning to count with cnn boosting. In *European Conference on Computer Vision*, pages 660–676, 2016.

[246] J. Wan and A. Chan. Adaptive density map generation for crowd counting. In *International Conference on Computer Vision*, pages 1130–1139, 2019.

[247] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, and R. Yang. Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12(1):5915, 2021.

[248] Su Wang. Generative Adversarial Networks (GAN): A Gentle Introduction. Technical report, Tutorial on GAN in LIN395C: Research in Computational Linguistics, University of Texas at Austin., 2017.

[249] Y. Wang and Y. Zou. Fast visual object counting via example-based density estimation. In *IEEE International Conference on Image Processing*, pages 3653–3657, 2016.

[250] D. A. Winkler. Use of artificial intelligence and machine learning for discovery of drugs for neglected tropical diseases. *Frontiers in Chemistry*, 9:614073, 2021.

[251] A. J. M. Wollman, R. Nudd, E. G. Hedlund, and M. C. Leake. From Animaculum to single molecules: 300 years of the light microscope. *Open Biology*, 5(4):150019, 2015.

[252] F. Wong, C. De La Fuente-Nunez, and J. J. Collins. Leveraging artificial intelligence in the fight against infectious diseases. *Science*, 381(6654):164–170, 2023.

[253] W. Xie, J. A. Noble, and A. Zisserman. Microscopy cell counting and detection with fully convolutional regression networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):283–292, 2018.

[254] Y. Xiong, X. Ba, A. Hou, K. Zhang, L. Chen, and T. Li. Automatic detection of mycobacterium tuberculosis using artificial intelligence. *Journal of Thoracic Disease*, 10(3):1936–1940, 2018.

[255] S. Yan, H. Liu, L. Sun, M. Zhou, Z. Xiao, and Q. Zhuang. Detection of Mycobacterium Tuberculosis in Ziehl-Neelsen Sputum Smear Images. In *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pages 1–6, 2018.

[256] B. Yesilkaya, M. Perc, and Y. Isler. Manifold learning methods for the diagnosis of ovarian cancer. *Journal of Computational Science*, 63:101775, 2022.

[257] H. Yousefi, F. Mohammadi, N. Mirian, and N. Amini. Tuberculosis bacilli identification: a novel feature extraction approach via statistical shape and color models. In *International Conference on Machine Learning and Applications*, pages 366–371, 2020.

[258] H. Yu, W. Jing, R. Iriya, Y. Yang, K. Syal, M. Mo, T. E. Grys, S. E. Haydel, S. Wang, and N. Tao. Phenotypic antimicrobial susceptibility testing with deep learning video microscopy. *Analytical Chemistry*, 90(10):6314–6322, 2018.

[259] H. Yu, P. Sun, F. He, and Z. Hu. A weighted region-based level set method for image segmentation with intensity inhomogeneity. *Plos One*, 16(8):e0255948, 2021.

[260] X. Yue, N. Dimitriou, and O. Arandjelovic. Colorectal cancer outcome prediction from H&E whole slide images using machine learning and automatically inferred phenotype profiles. In *International Conference on Bioinformatics and Computational Biology*, pages 139–149, 2019.

[261] S. Yuheng and Y. Hao. Image segmentation algorithms overview. In *Asia Modelling Symposium*, pages 103–107, 2017.

[262] M. Zachariou, O. Arandjelović, E. Dombay, W. Sabiiti, B. Mtafya, N. E. Ntinginya, and D. J. Sloan. Localization and phenotyping of tuberculosis bacteria using a combination of deep learning and SVMs. *Computers in Biology and Medicine*, 167:107573, 2023.

[263] M. Zachariou, O. Arandjelović, E. Dombay, W. Sabiiti, B. Mtafya, and D. Sloan. Extracting and Classifying Salient Fields of View From Microscopy Slides of Tuberculosis Bacteria. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 146–157, 2022.

[264] M. Zachariou, O. Arandjelović, E. Dombay, W. Sabiiti, B. Mtafya, and J. D. Sloan. Estimating Phenotypic Characteristics of Tuberculosis Bacteria. In *Symposium on Applied Computing*, pages 1110–1113, 2023.

[265] M. Zachariou, O. Arandjelović, W. Sabiiti, B. Mtafya, and D. Sloan. Tuberculosis bacteria detection and counting in fluorescence microscopy images using a multi-stage deep learning pipeline. *Information*, 13(2):96, 2022.

[266] M. Zachariou, O. Arandjelović, and D. J. Sloan. Automated methods for tuberculosis detection/diagnosis: A literature review. *BioMedInformatics*, 3(3):724–751, 2023.

[267] M. Zachariou, N. Dimitriou, and O. Arandjelović. Visual reconstruction of ancient coins using cycle-consistent generative adversarial networks. *Sci*, 2(3):52, 2020.

[268] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[269] T. Zahir, R. Camacho, R. Vitale, C. Ruckebusch, J. Hofkens, M. Fauvart, and J. Michiels. High-throughput time-resolved morphology screening in bacteria reveals phenotypic responses to antibiotics. *Communications Biology*, 2(1):1–13, 2019.

[270] Y. Zhai, Y. Liu, D. Zhou, and S. Liu. Automatic identification of mycobacterium tuberculosis from ZN-stained sputum smear: Algorithm and system design. In *International Conference on Robotics and Biomimetics*, pages 41–46, 2010.

[271] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, L. Ye, M. Gao, Z. Zhou, L. Li, J. Wang, Z. Yang, H. Cai, J. Xu, L. Yang, W. Cai, W. Xu, S. Wu, W. Zhang, S. Jiang, L. Zheng, X. Zhang, L. Wang, L. Lu, J. Li, H. Yin, W. Wang, O. Li, C. Zhang, L. Liang, T. Wu, R. Deng, K. Wei, Y. Zhou, T. Chen, J. Yiu-Nam Lau, M. Fok, J. He, T. Lin, W. Li, and G. Wang. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. *Cell*, 182(5):1360, 2020.

[272] Y. Zhang, W. W. Yew, and M. R. Barer. Targeting persisters for tuberculosis control. *Antimicrobial Agents and Chemotherapy*, 56(5):2223–2230, 2012.

[273] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision*, pages 2223–2232, 2017.

[274] J. Zhuang, T. Tang, Y. Ding, S. Tatikonda, N. Dvornek, X. Papademetris, and J. S. Duncan. Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in Neural Information Processing Systems*, 33:18795–18806, 2020.

[275] Y. Zou, H. Bu, L. Guo, Y. Liu, J. He, and X. Feng. Staining with two observational methods for the diagnosis of tuberculous meningitis. *Experimental and Therapeutic Medicine*, 12(6):3934–3940, 2016.

Appendices

Appendix A

Chapter 3

A.1 Staining and microscopy procedures

A.1.1 Preparation of staining dyes

A.1.1.1 HCS LipidTOX™ Red neutral lipid stain

HCS LipidTOX™ Red Neutral Lipid Stain (Thermo Fisher Scientific, US) is diluted in a ratio of 1:500 using sterile Phosphate buffer solution (PBS). For example, 2 μL of the dye is mixed with 998 μL of PBS. Then, 500 μL of the 1:500 LTR solution is applied to a smear for each slide.

A.1.1.2 Auramine O staining solution

Ready-made Auramine O solution (Sigma-Aldrich, UK) is used. 100 μL of the solution to cover the smear is applied for each slide.

A.1.1.3 Potassium permanganate

To prepare a 0.1% solution, one should dissolve 0.1 g of potassium permanganate powder (Sigma-Aldrich, UK) into 100 mL of sterile distilled water. This solution should be stored at room temperature. For each slide smear, 500 μL of the solution should be used.

A.1.2 Preparation of sputum smears

- A total of 100 μL of the NALC-NaOH decontaminated sputum pellet should be applied to each slide.

- The smears require air drying for 30 minutes at room temperature.

- Slide fixation should occur on a hotplate set at 80°C for a duration of 20 minutes.
- The slides need to be fixed with a 20.9% (w/v) formalin solution overnight, followed by air drying the next day to remove excess formalin.
- Heat and formalin fixed slides should be stored in a slide box at room temperature (20-23°C) until required for Auramine/LTR staining and microscopy.

A.1.3 Auramine and LipidTox Red staining procedure

- The formalin-fixed slides should be covered with 100 μL of Auramine O solution and allowed to incubate for 15 minutes at room temperature, shielded from light.
- The slides ought to be washed twice with 1 mL of distilled water and then decolorized with 500 μL of 0.5% acid alcohol for 1 minute.
- A subsequent washing of the slides should occur twice with 1 mL of distilled water.
- The slides should then be covered with 500 μL of LTR for 20 minutes at room temperature, with protection from light using aluminum foil.
- A second washing should be performed twice with 1 mL of distilled water.
- The slides require coverage with 500 μL of 0.5% potassium permanganate solution for 45 seconds to bleach the non-specific background, followed by two washes with 1 mL of distilled water.
- Smears should then be mounted with 30 μL of PBS and covered with a coverslip.
- Transparent nail polish may be applied at the edges of the coverslip to secure it in place during microscopic examination.

A.1.4 Fluorescence microscopy

- The stained slides should be covered with aluminum foil to shield them from light before examination using a Leica DM5500 microscope (Leica).
- An examination should be conducted using a 100× oil immersion objective lens, with the immersion oil being specific to Leica microscopy (Leica immersion oil, ISO8036).

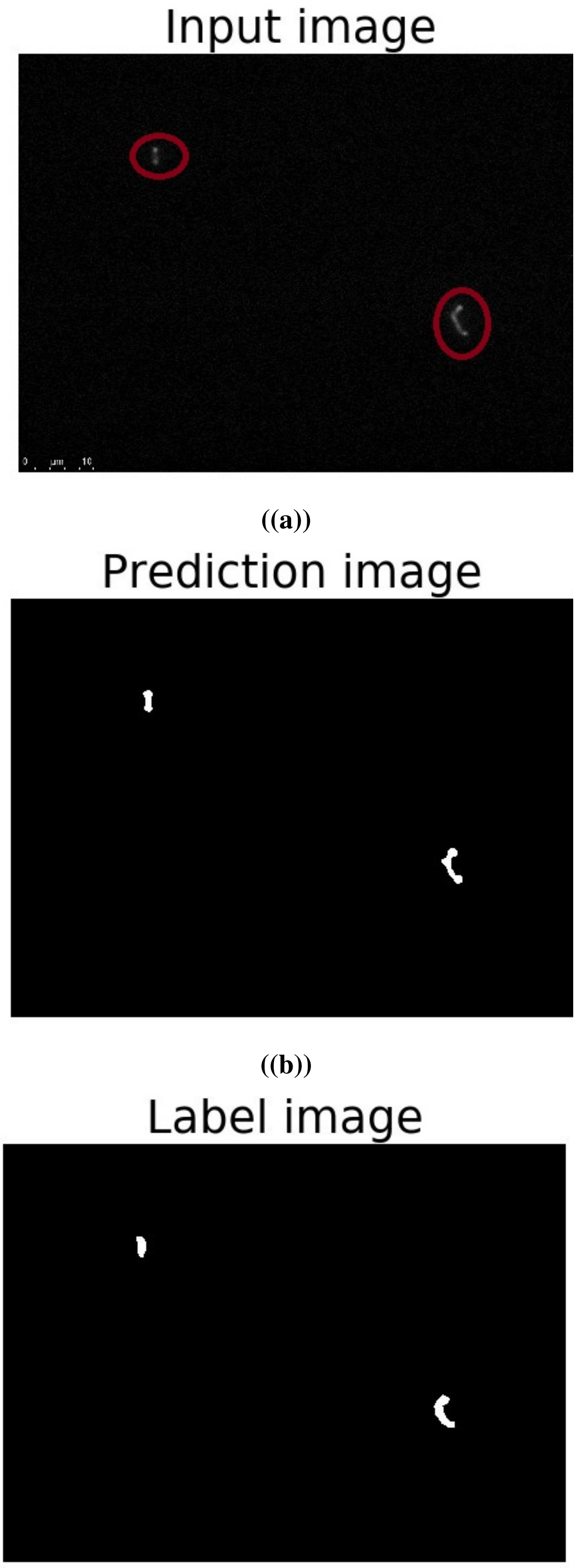
- Image capture is to be performed using the Leica camera DFC 3000 G.
- The systematic scanning of slides should encompass three clean sweeps.
- Each FOV that contains potential *Mtb* cells must be examined using two filter cubes:
  - An N3 filter cube, with excitation and emission spectra of 546/12 and 600/40nm, should be used to observe Auramine O stained *Mtb* cells.
  - A TX2 filter cube, with excitation and emission spectra of 560/40 and 645/75nm, should be used to observe intracellular lipids.

A.2 Chapter 3 segmentation cases examples

Included below are visual illustrations representing the three distinct scenarios outlined in Table 4.1. The initial section will showcase images falling within the first scenario, characterized by clearly defined bacteria. Following sections will present examples from the second and third scenarios, respectively. Moreover, the figures are presented in a triad format, consisting of the original image, the predicted image, and the labelled image. This arrangement provides supplementary proof of the method's performance.

A.2.1 Case 1 images

The subsequent figures serve to illustrate examples from case 1 and demonstrate the method's approach to handling such instances. The bacteria responsible for the images depicted in this section have been encircled in red.

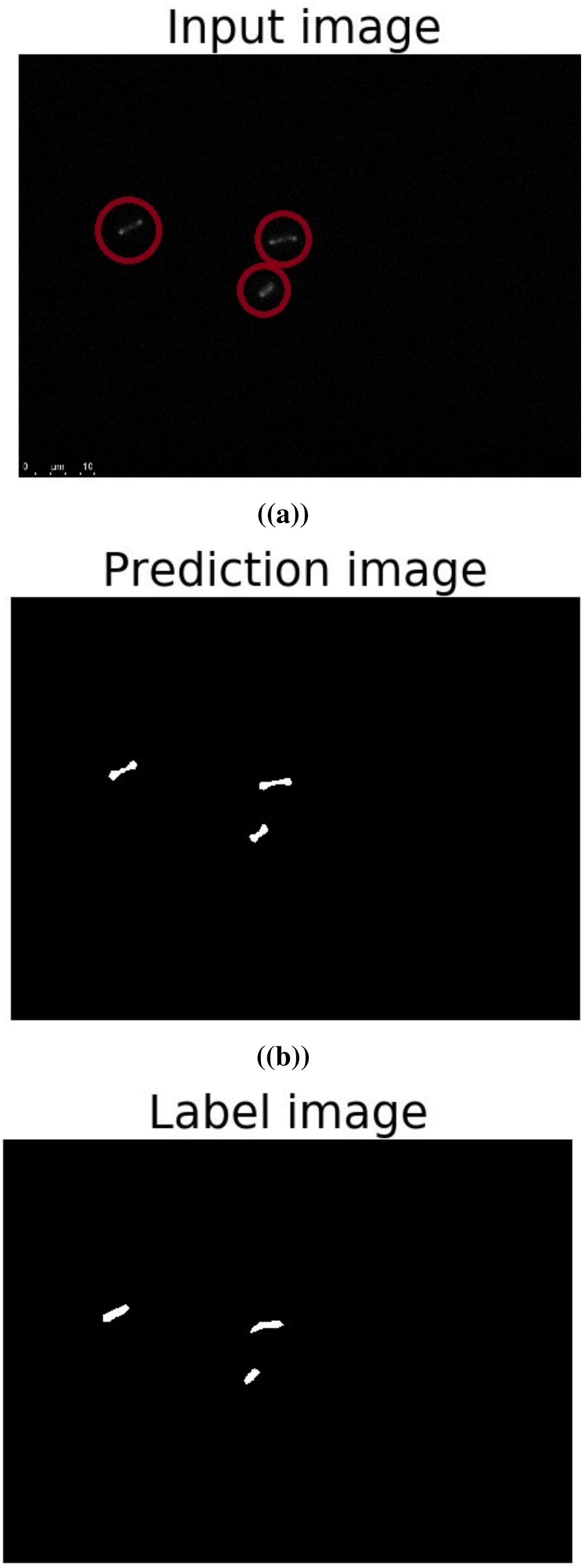


((a))

((b))

((c))

Figure A.1: Case number 1 relates to objects that exhibit clear visibility with distinct pixel intensity and sharpness, consistent with stained *Mtb* cells

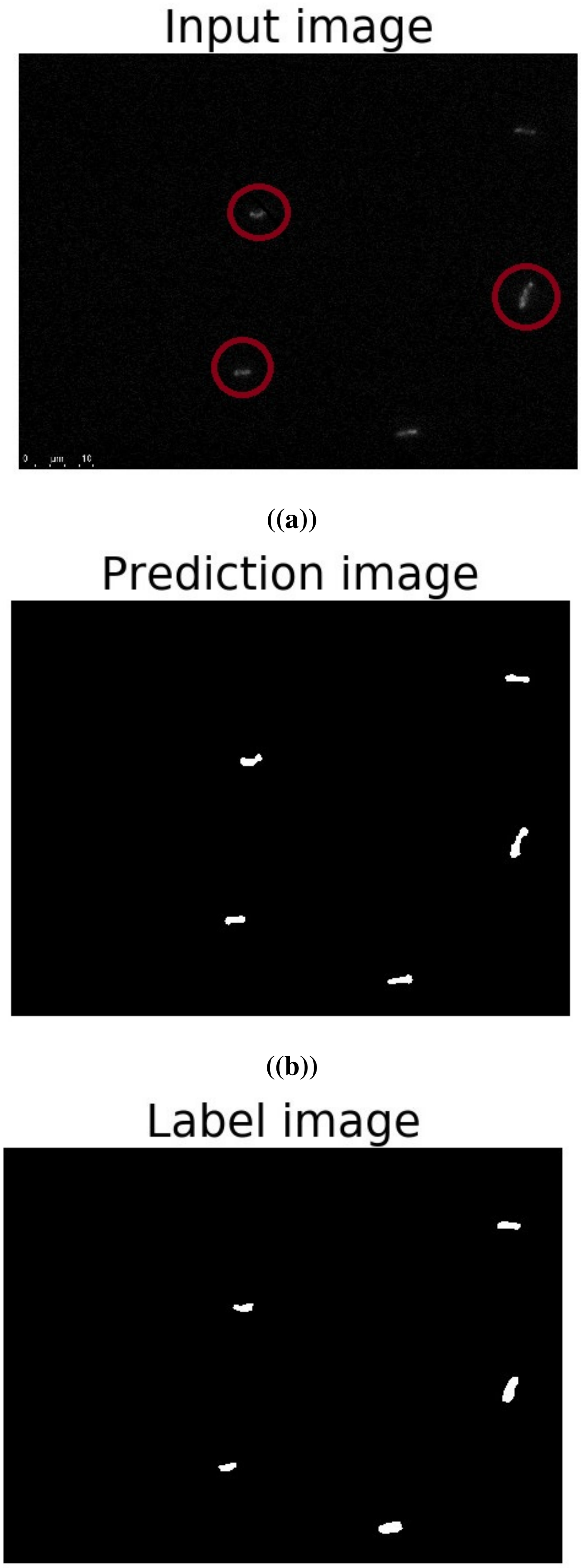


((a))

((b))

((c))

Figure A.2: Case number 1 relates to objects that exhibit clear visibility with distinct pixel intensity and sharpness, consistent with stained *Mtb* cells



((a))

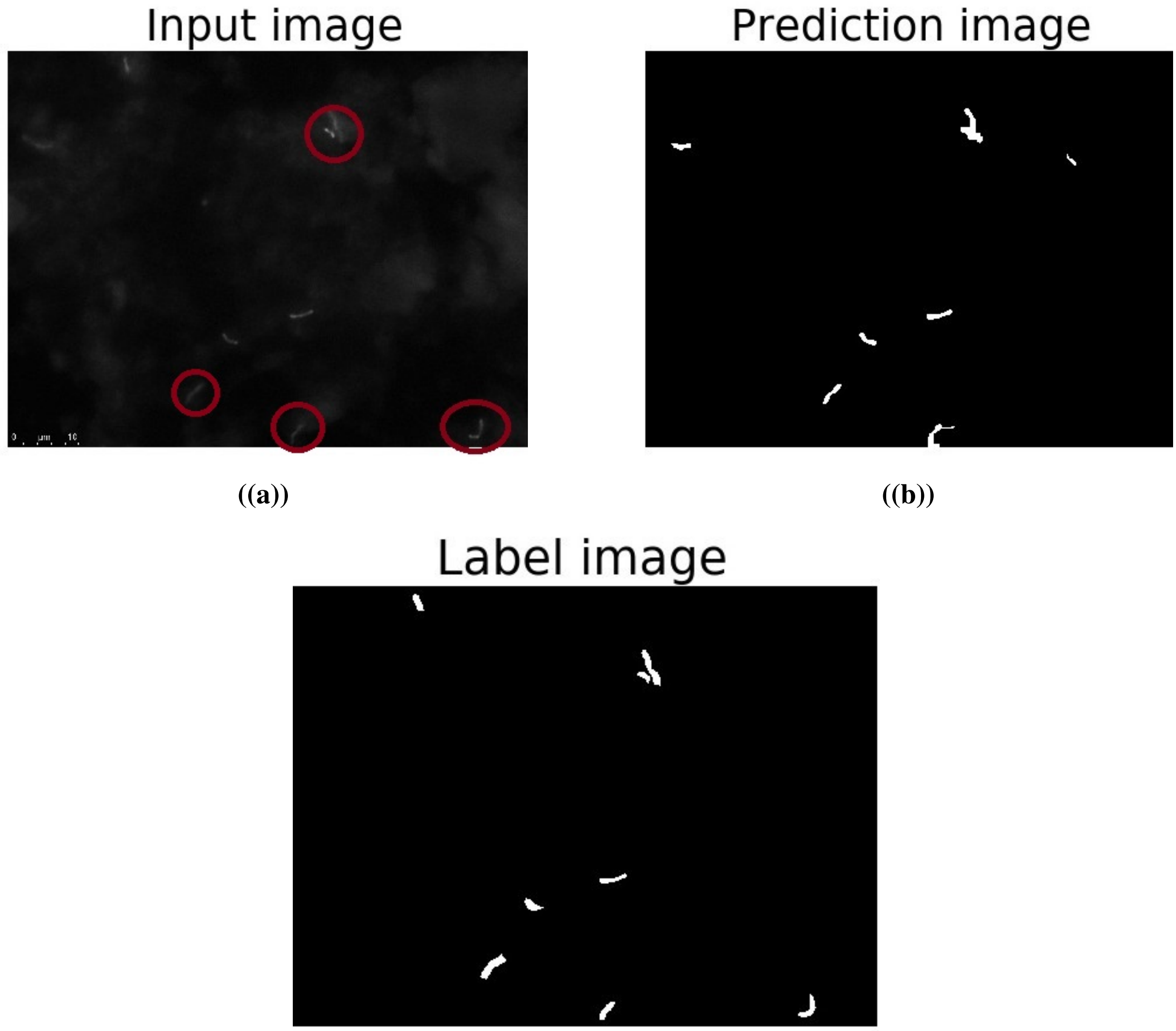
((b))

((c))

Figure A.3: Case number 1 relates to objects that exhibit clear visibility with distinct pixel intensity and sharpness, consistent with stained *Mtb* cells

A.2.2 Case 2 images

The subsequent figures serve to illustrate examples from case 2 and demonstrate the method's approach to handling such instances. The bacteria responsible for the images depicted in this section have been encircled in red.

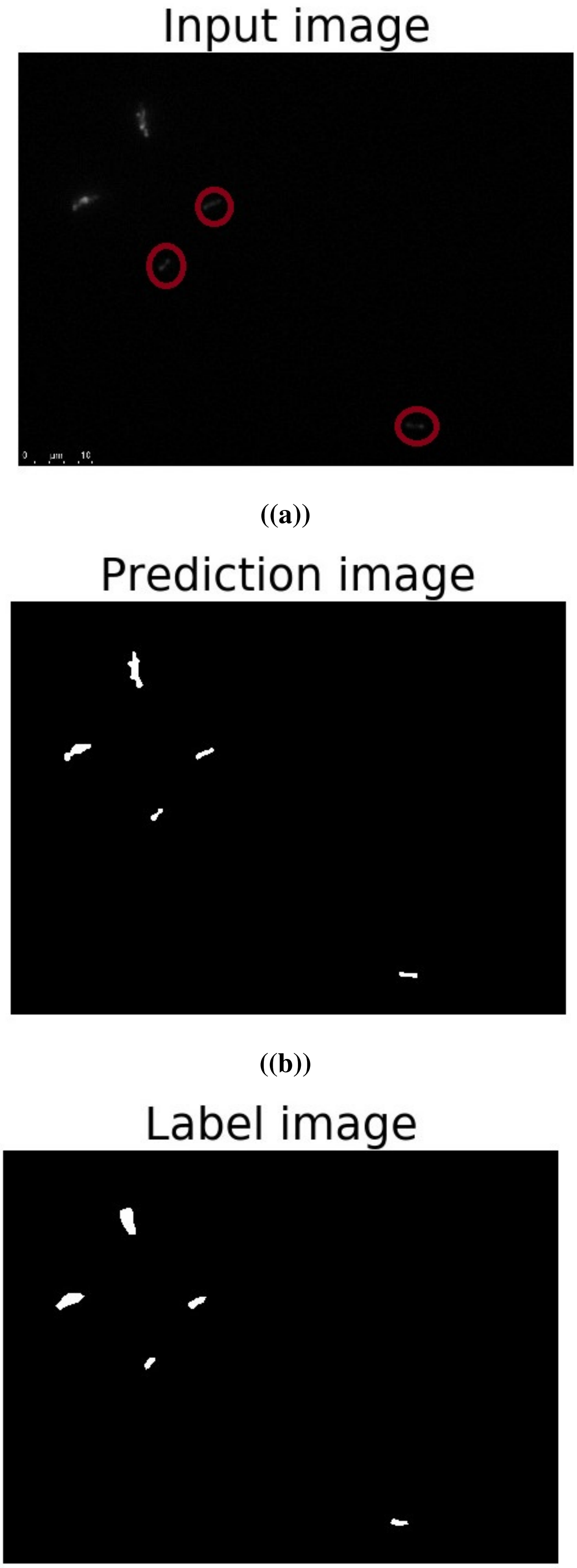


((a))

((b))

((c))

Figure A.4: Case number 2 relates to objects with diminished pixel intensity and reduced clarity, but which remain distinguishable from the background as possible *Mtb* cells.

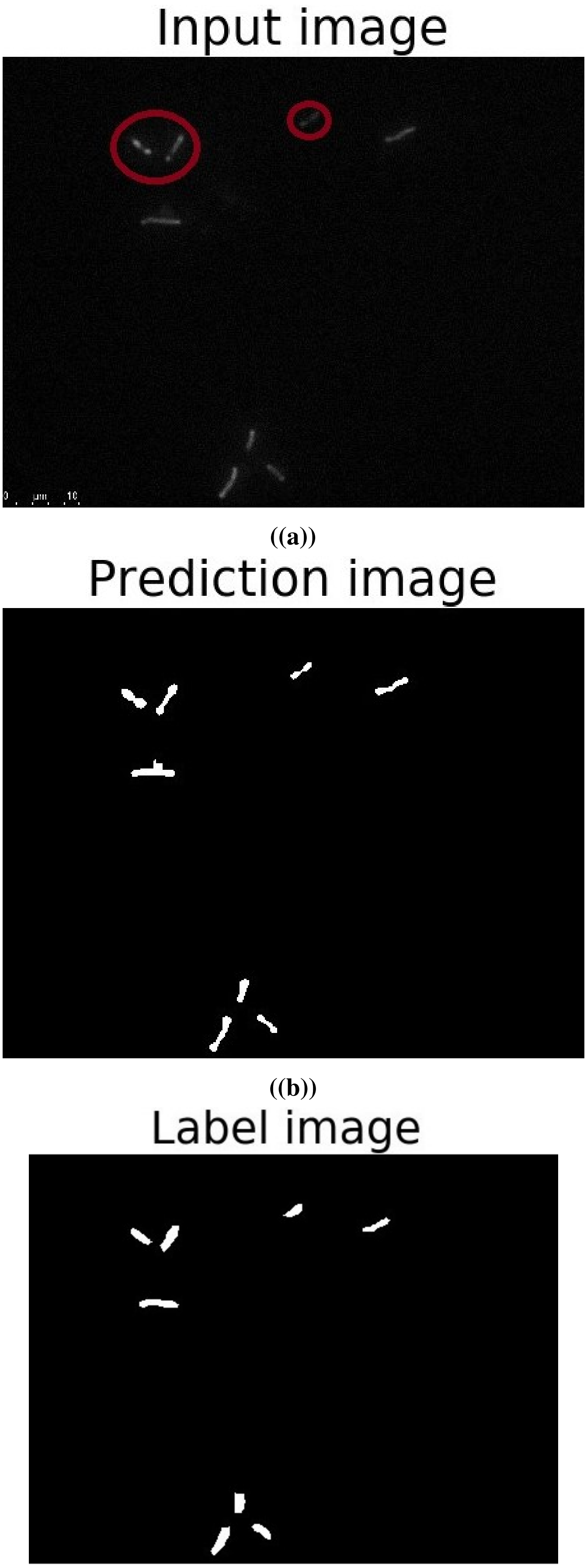


((a))

((b))

((c))

Figure A.5: Case number 2 relates to objects with diminished pixel intensity and reduced clarity, but which remain distinguishable from the background as possible *Mtb* cells.



((a))

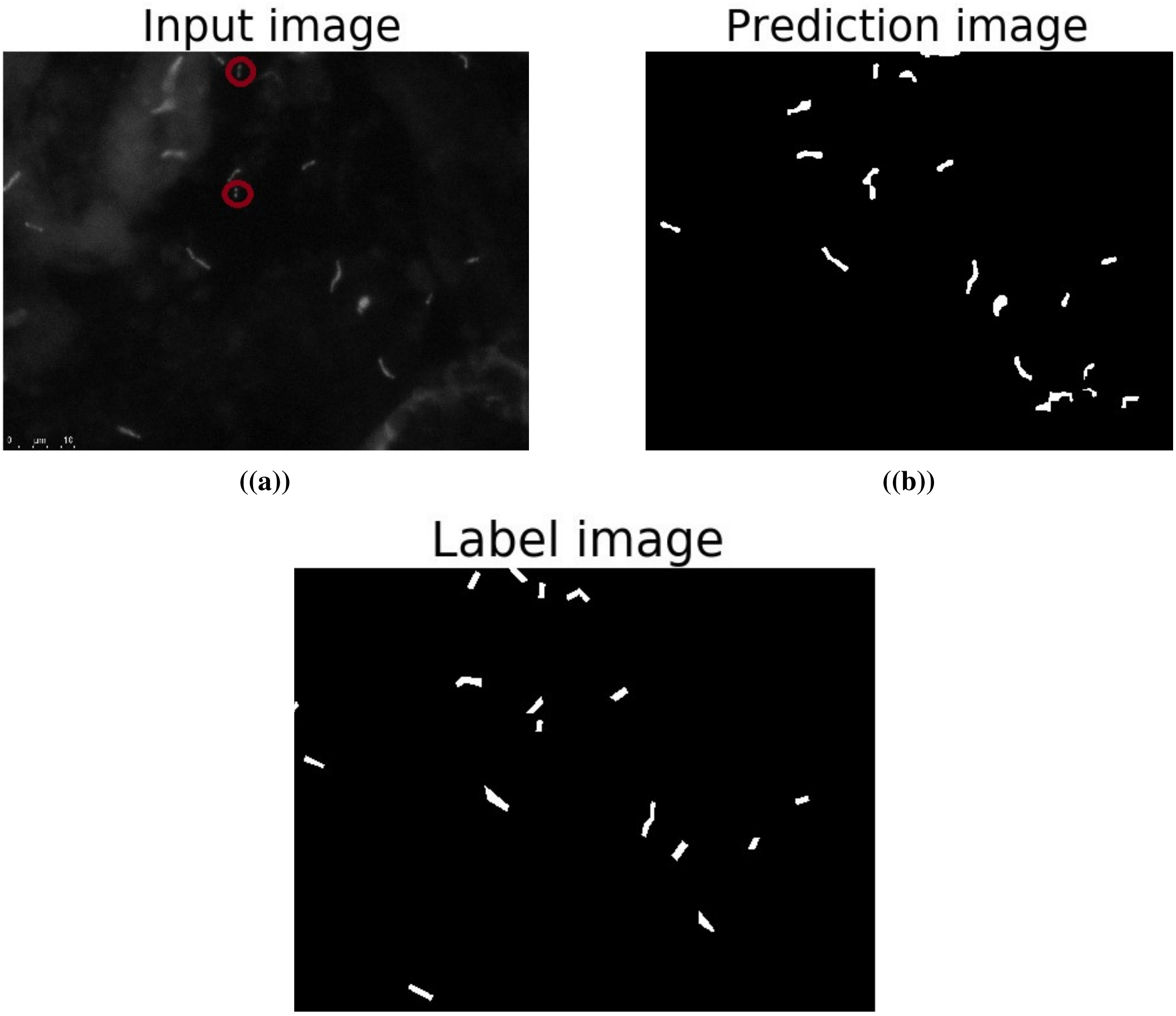
((b))

((c))

Figure A.6: Case number 2 relates to objects with diminished pixel intensity and reduced clarity, but which remain distinguishable from the background as possible *Mtb* cells.

A.2.3 Case 3 images

The subsequent figures serve to illustrate examples from case 3 and demonstrate the method's approach to handling such instances. The bacteria responsible for the images depicted in this section have been encircled in red.


Input image

Prediction image

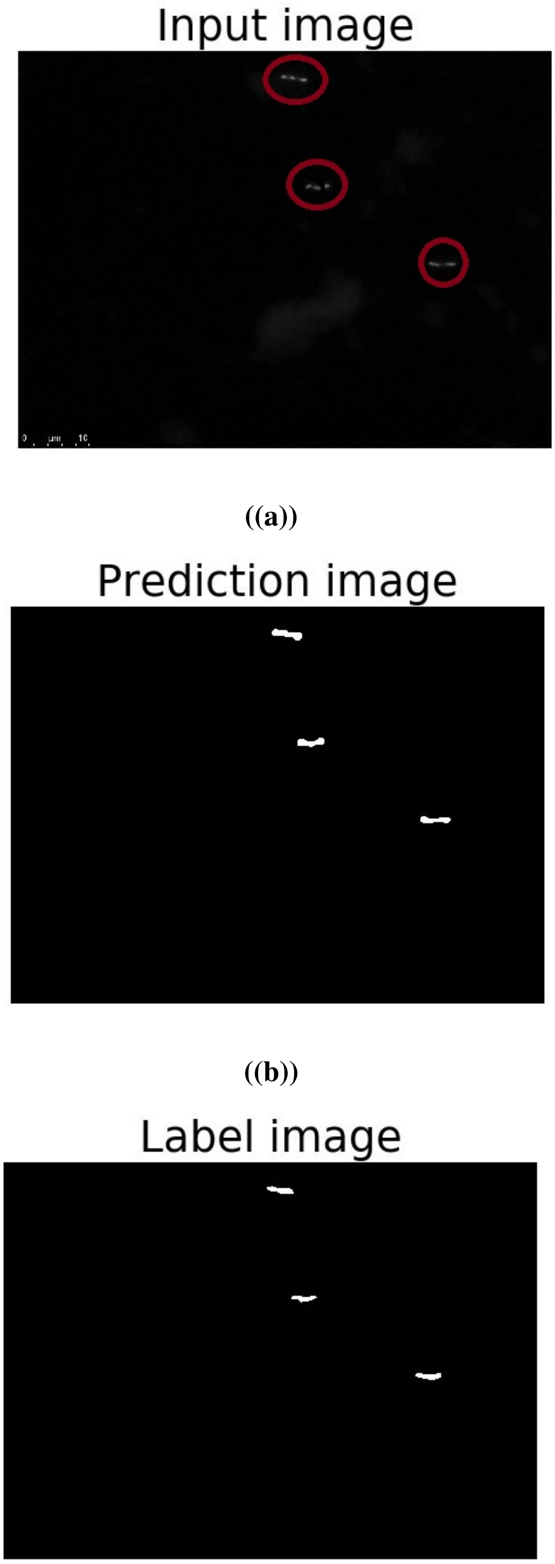
Label image

((a))

((b))

((c))

Figure A.7: Case number 3 relates case, is for objects with inconsistent pixel intensity in specific areas of their shape (perhaps due to variable fluorescent dye update) but which, overall, have features that are still compatible with *Mtb* cell morphology

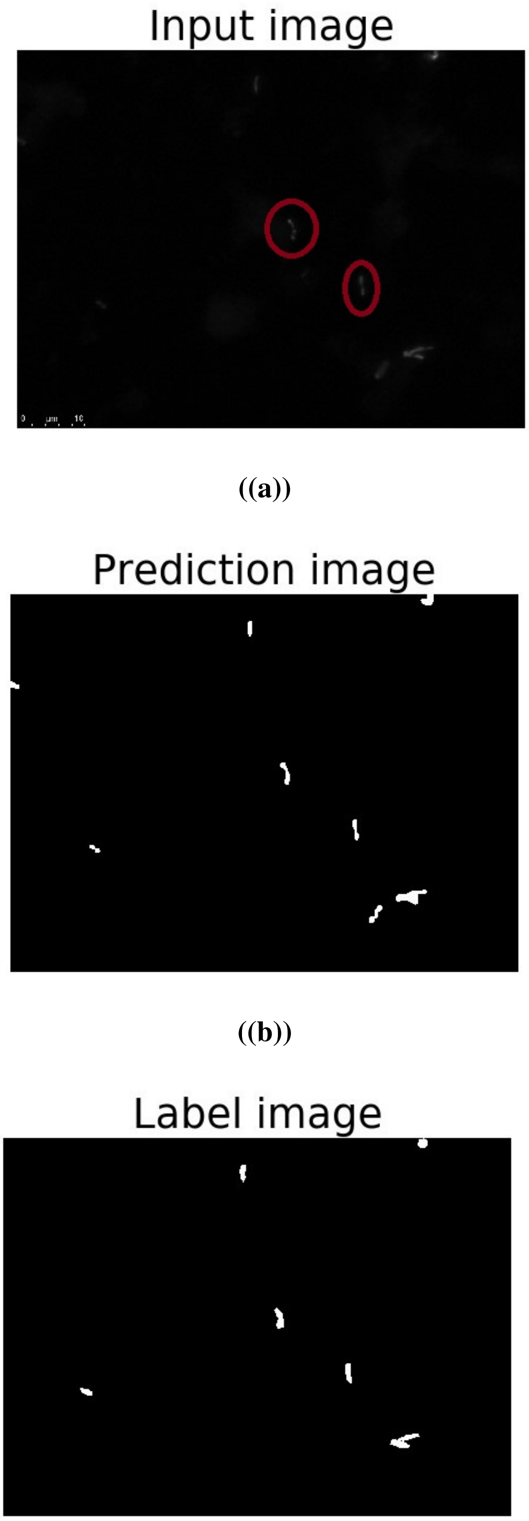


((a))

((b))

((c))

Figure A.8: Case number 3 relates case, is for objects with inconsistent pixel intensity in specific areas of their shape (perhaps due to variable fluorescent dye update) but which, overall, have features that are still compatible with *Mtb* cell morphology



((a))

((b))

((c))

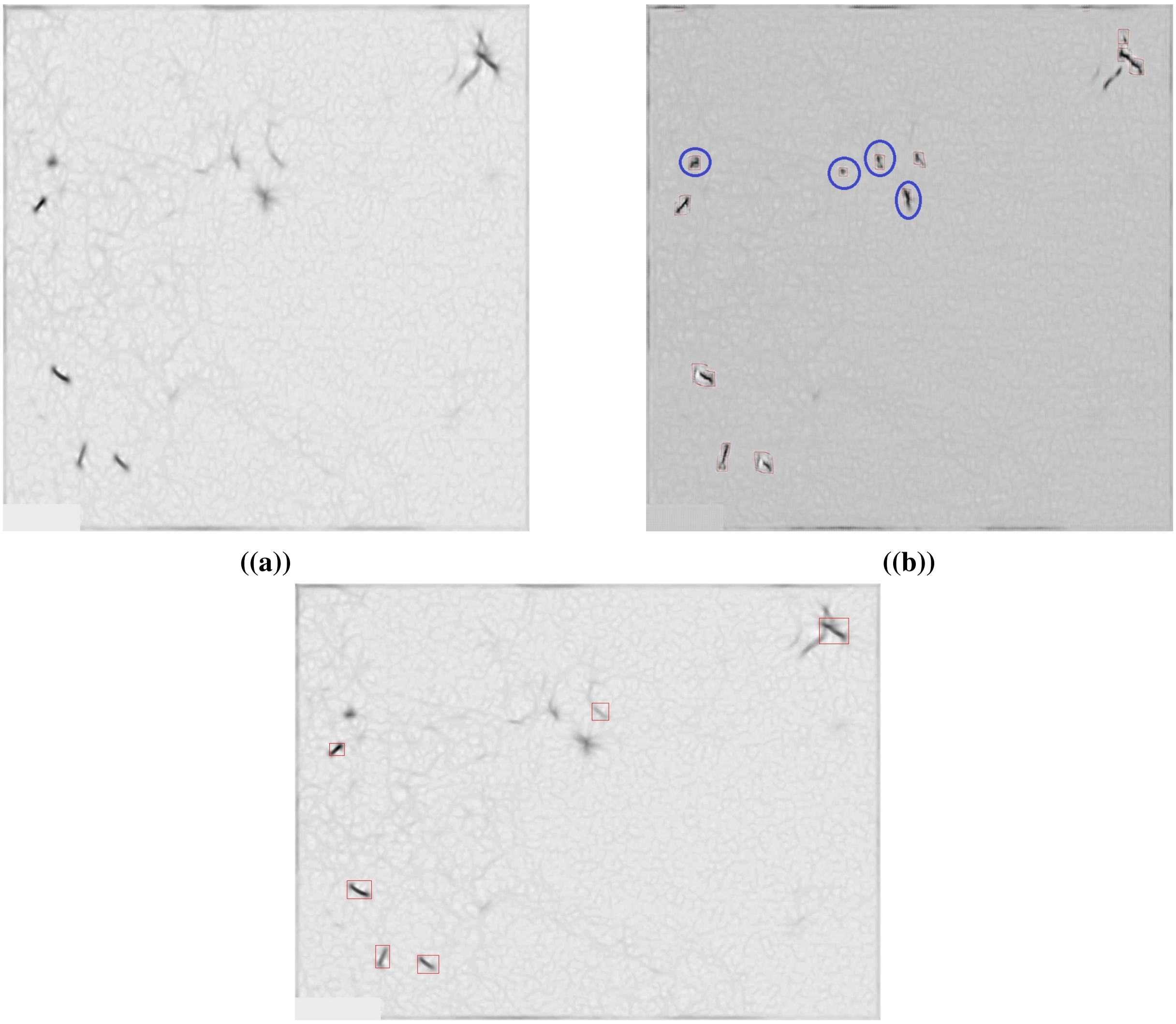
Figure A.9: Case number 3 relates case, is for objects with inconsistent pixel intensity in specific areas of their shape (perhaps due to variable fluorescent dye update) but which, overall, have features that are still compatible with *Mtb* cell morphology

Appendix B

Chapter 4

B.1 Cycle-GANs synthetic images

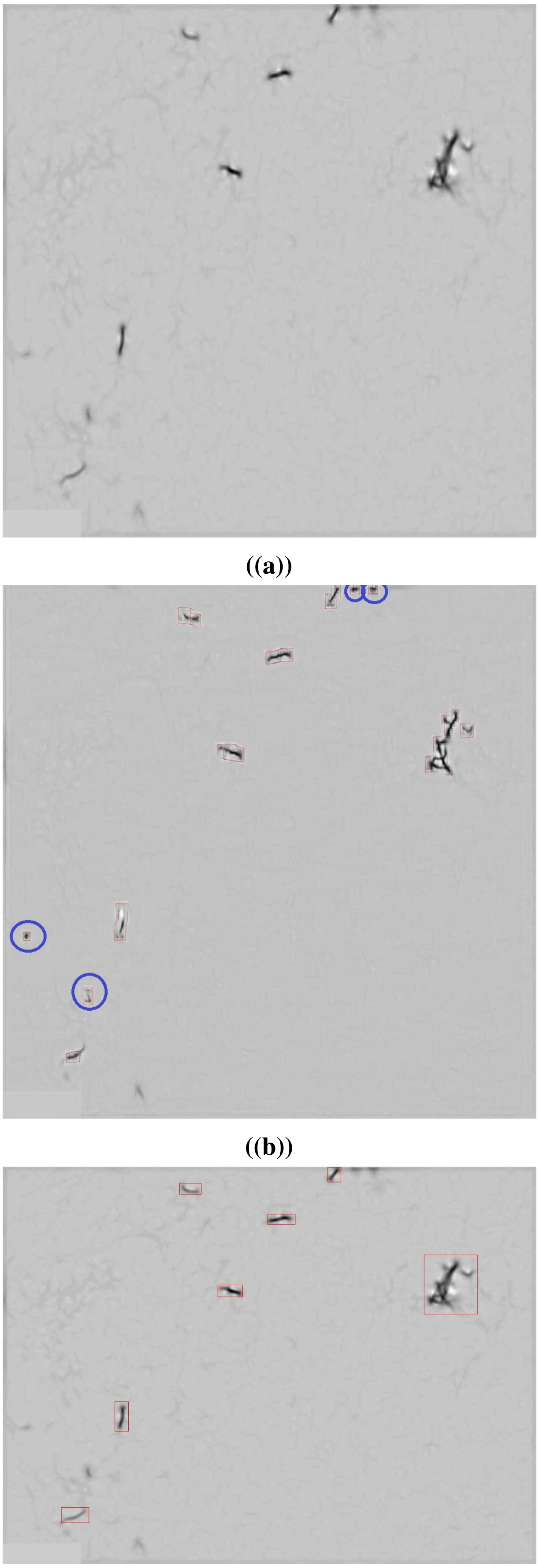
In this section, I present the outcomes of the Cycle-GANs, which occasionally misinterpret pixel fluctuations as bacteria, leading to inaccuracies in the translation of bacteria-like objects, including their bounding boxes. Each subsequent figure follows the format of a real (unlabeled image), a synthesized image, and a ground truth image.



((a)) ((b))

((c))

Figure B.1: Case number 1 of synthetic images. Incorrectly synthesised bacteria-like objects are circled in blue.

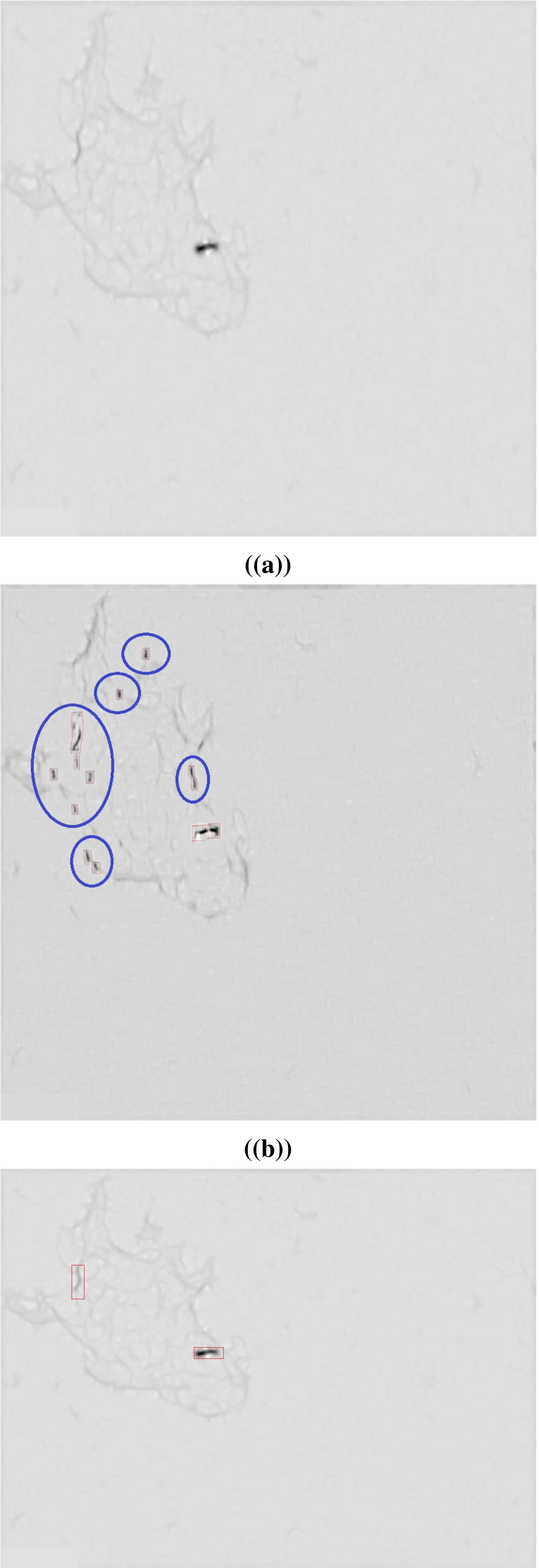


((a))

((b))

((c))

Figure B.2: Case number 2 of synthetic images. Incorrectly synthesised bacteria-like objects are circled in blue.

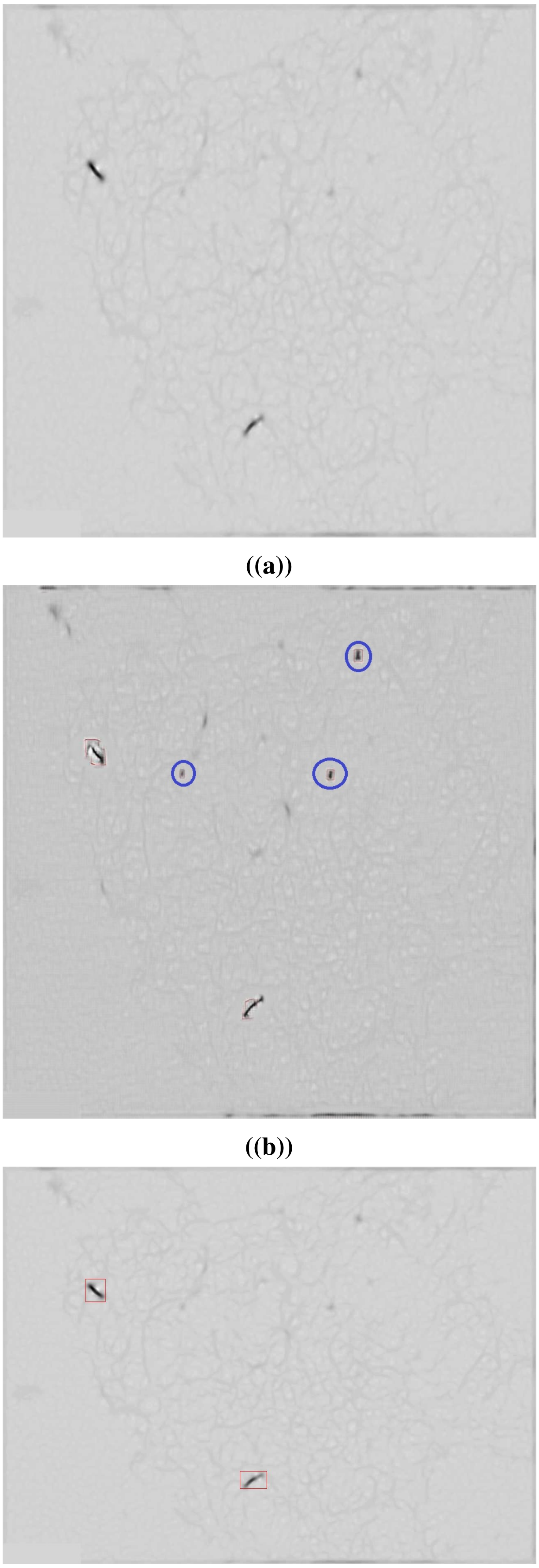


((a))

((b))

((c))

Figure B.3: Case number 3 of synthetic images. Incorrectly synthesised bacteria-like objects are circled in blue.

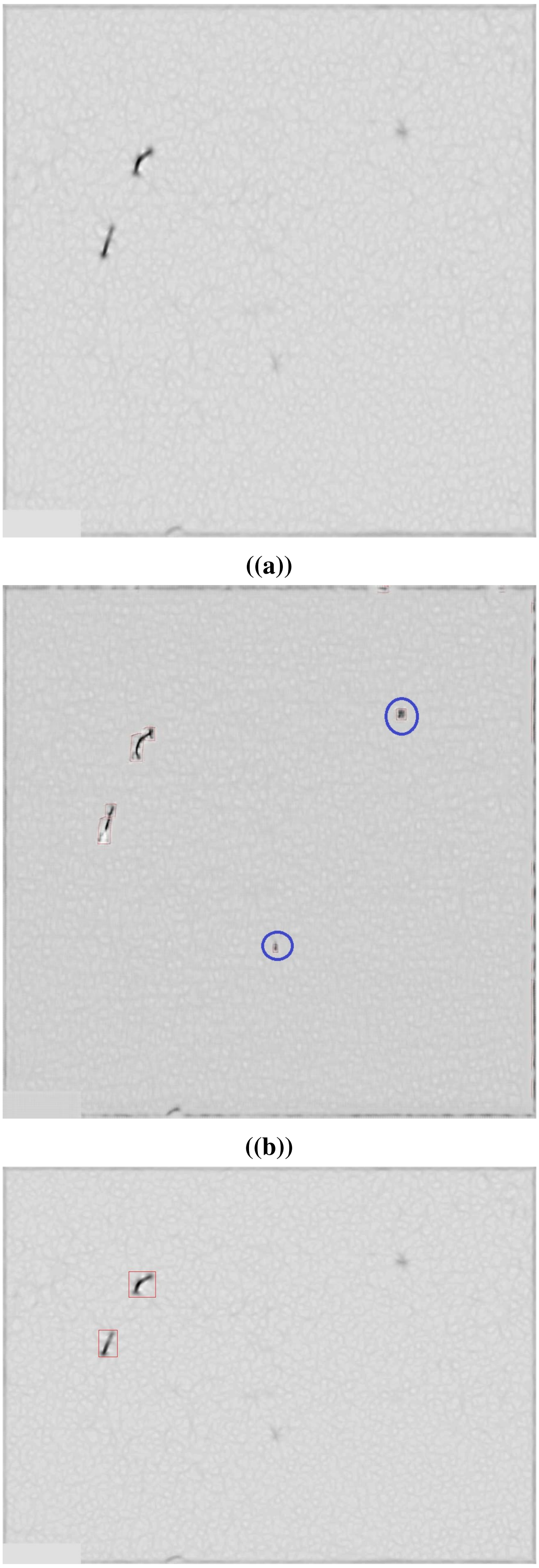


((a))

((b))

((c))

Figure B.4: Case number 4 of synthetic images. Incorrectly synthesised bacteria-like objects are circled in blue.

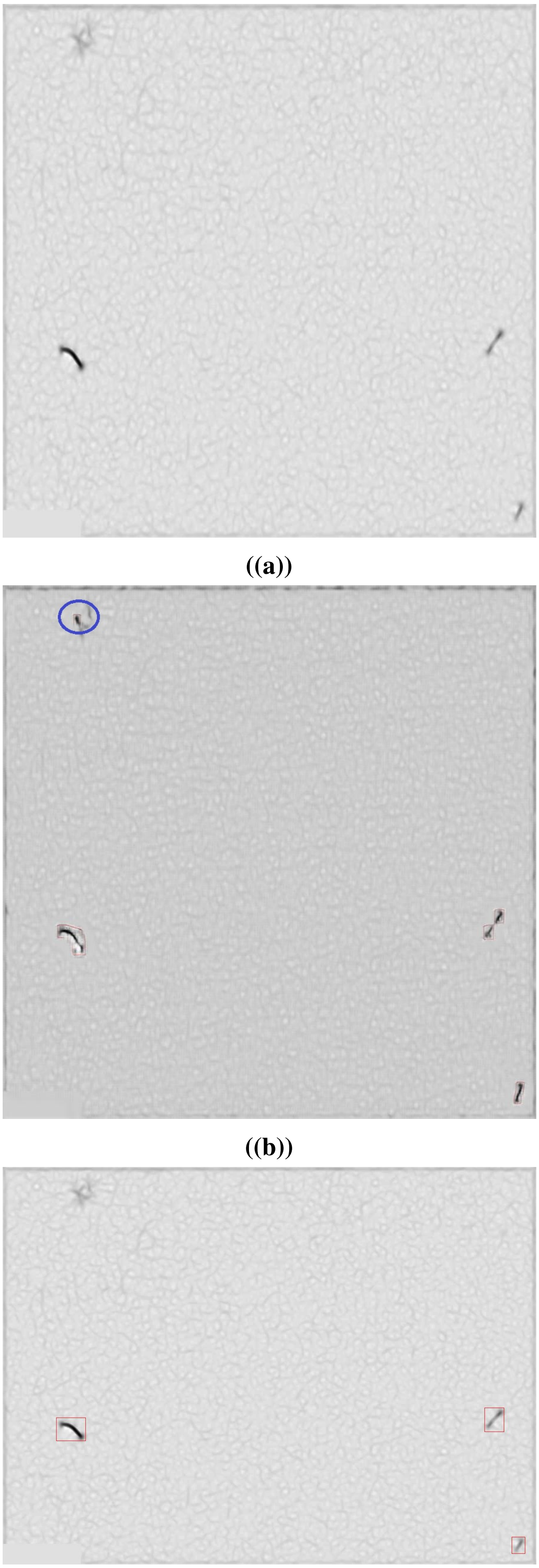


((a))

((b))

((c))

Figure B.5: Case number 5 of synthetic images. Incorrectly synthesised bacteria-like objects are circled in blue.

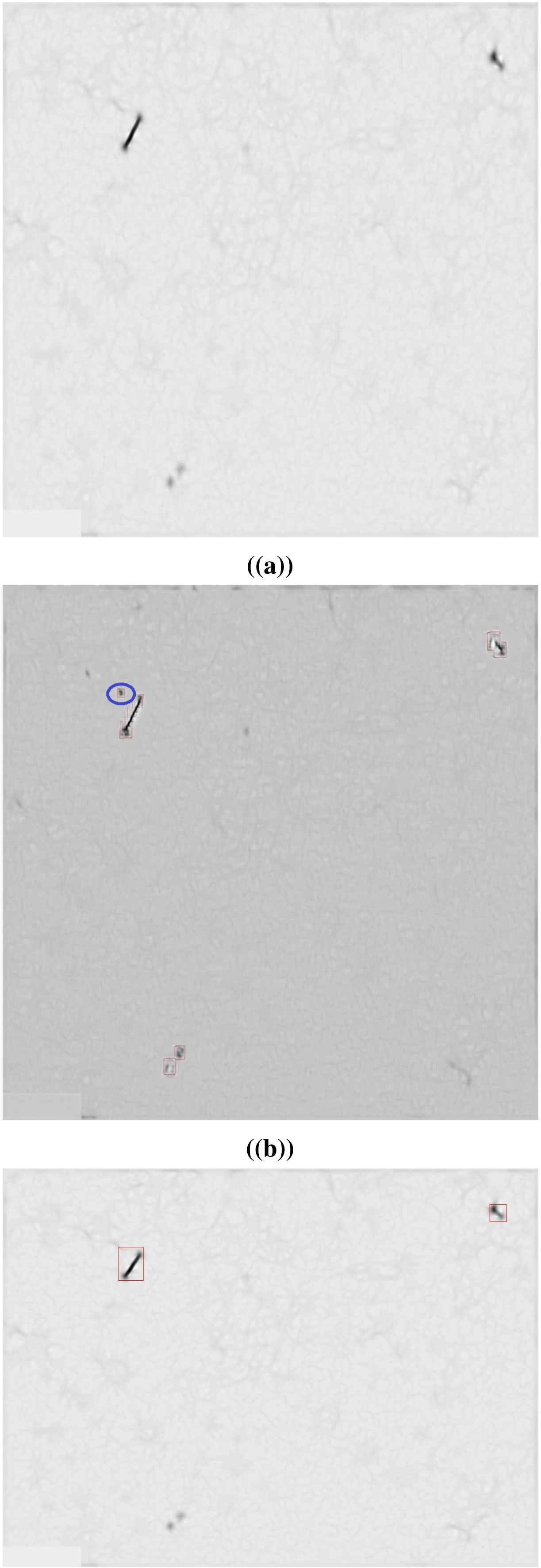


((a))

((b))

((c))

Figure B.6: Case number 6 of synthetic images. Incorrectly synthesised bacteria-like objects are circled in blue.



((a))

((b))

((c))

Figure B.7: Case number 7 of synthetic images. Incorrectly synthesised bacteria-like objects are circled in blue.

Appendix C

Chapter 6

C.1 Proposed model for semantic segmentation

In this appendix section, I provide additional results from the model introduced in Chapter 6, pertaining to the semantic segmentation phase. I have organized the exemplar cases into two subsections: Auramine O FOVs and LTR FOVs.

C.1.1 Auramine O segmented FOVs

Below follow examples of segmented Auramine O FOVs.

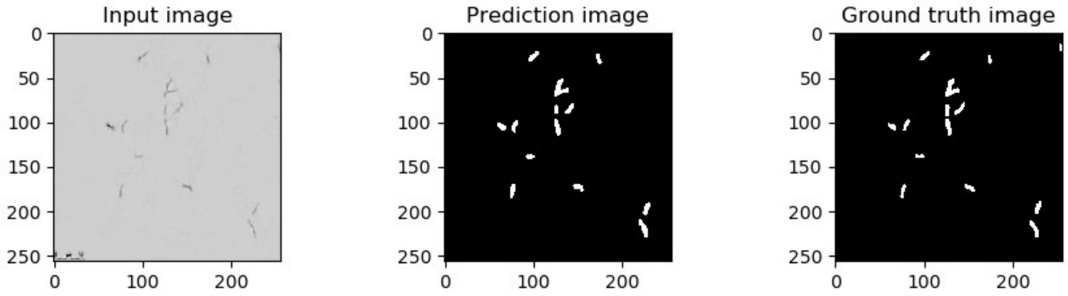


Figure C.1: Case number 1 of segmented images of Auramine O FOVs

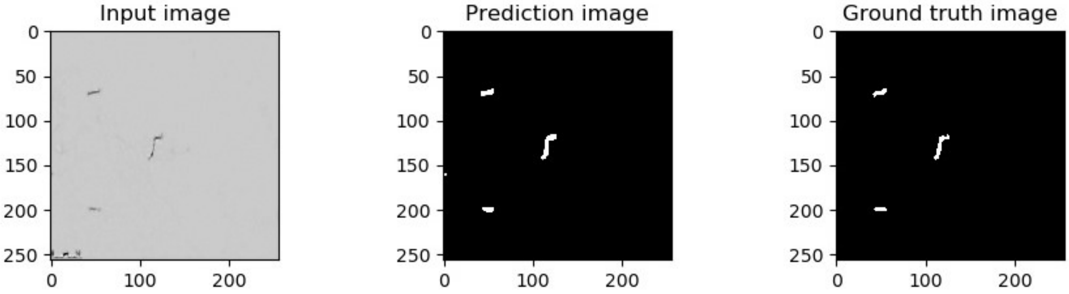


Figure C.2: Case number 2 of segmented images of Auramine O FOVs

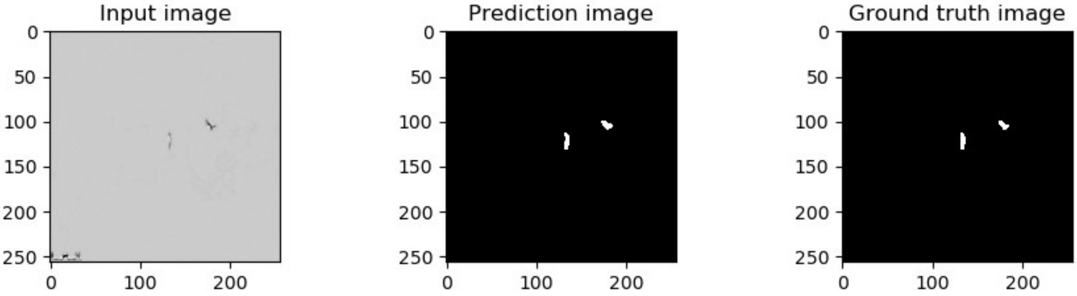


Figure C.3: Case number 3 of segmented images of Auramine O FOVs

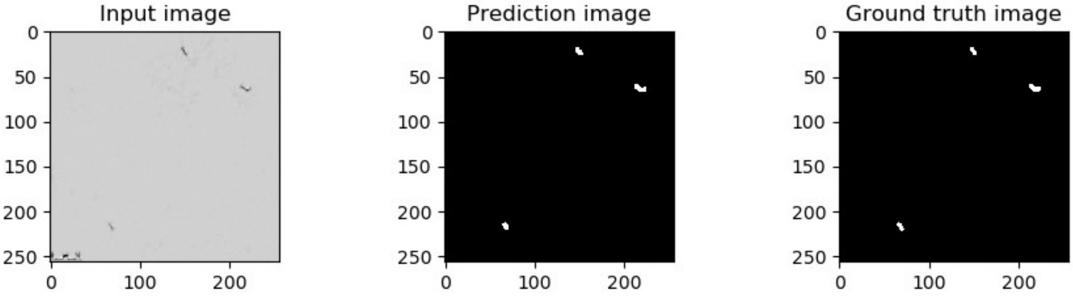


Figure C.4: Case number 4 of segmented images of Auramine O FOVs

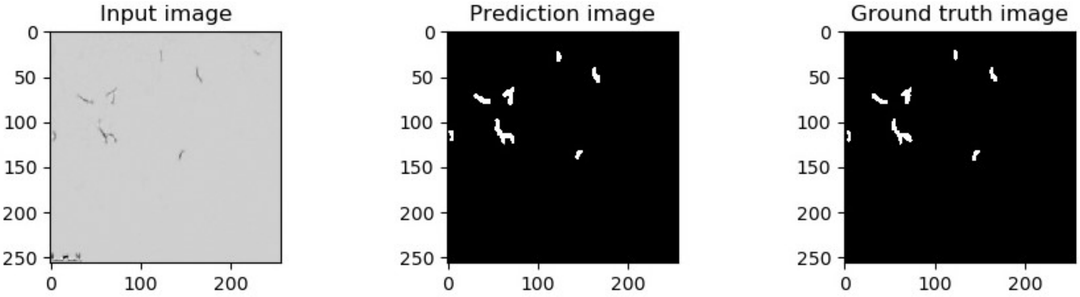


Figure C.5: Case number 5 of segmented images of Auramine O FOVs

C.1.2 LTR segmented FOVs

Below follow examples of segmented LTR FOVs.

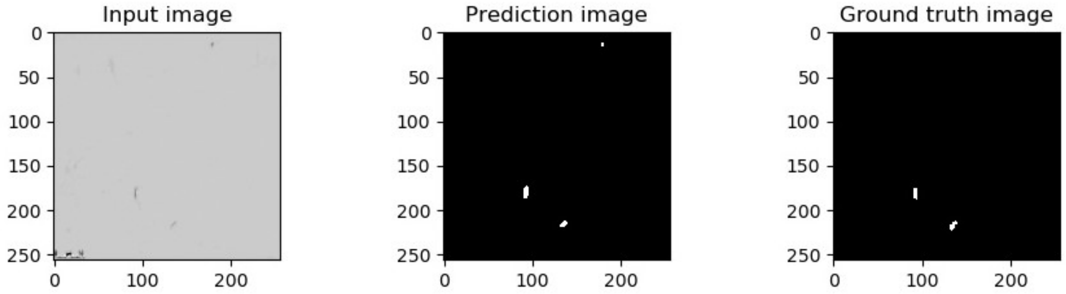


Figure C.6: Case number 1 of segmented images of LTR FOVs

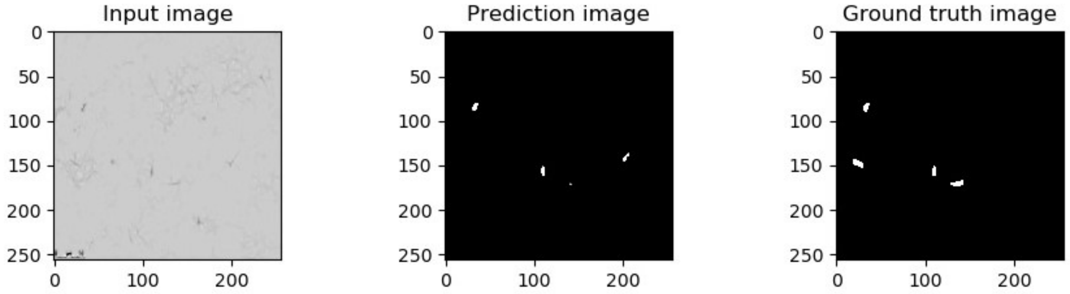


Figure C.7: Case number 2 of segmented images of LTR FOVs

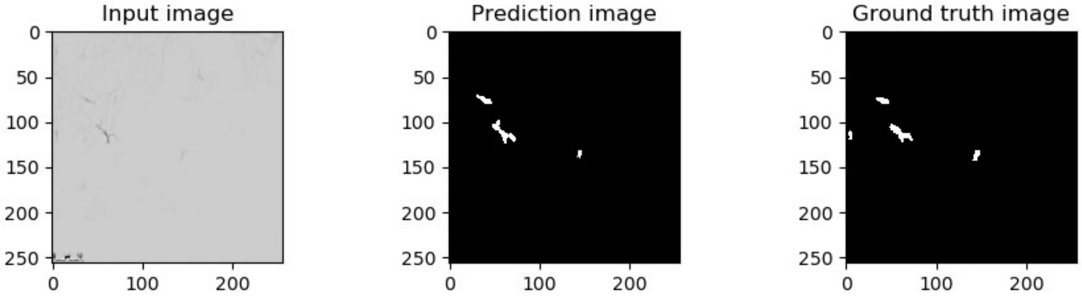


Figure C.8: Case number 3 of segmented images of LTR FOVs

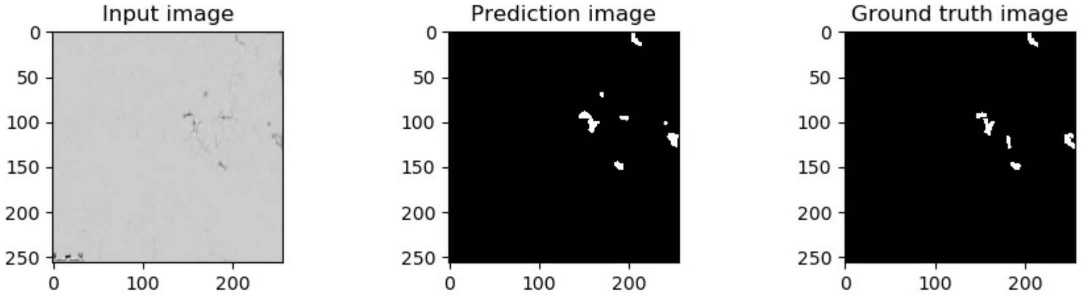


Figure C.9: Case number 4 of segmented images of LTR FOVs

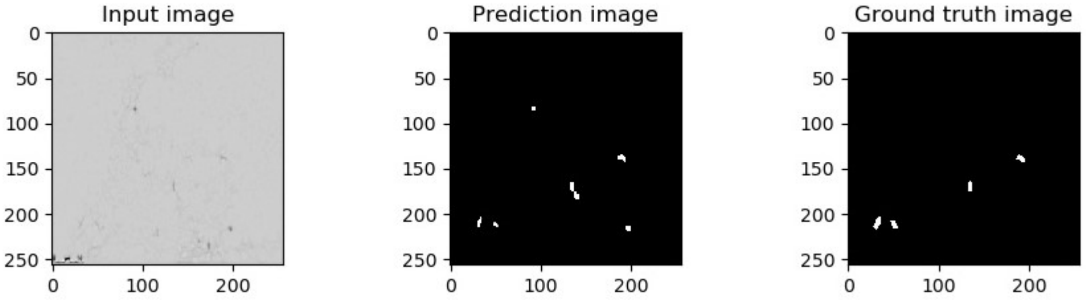


Figure C.10: Case number 5 of segmented images of LTR FOVs