BioMedInformatics

Review

Automated Methods for Tuberculosis Detection/Diagnosis: A Literature Review

Marios Zachariou [1,*], Ognjen Arandjelović [1] and Derek James Sloan [2]

[1] School of Computer Science, University of St Andrews, St Andrews KY16 9SX, UK; ognjen.arandjelovic@gmail.com
[2] School of Medicine, University of St Andrews, St Andrews KY16 9AJ, UK; djs26@st-andrews.ac.uk
* Correspondence: marios.zachariou@hotmail.com

Abstract: Tuberculosis (TB) is one of the leading infectious causes of death worldwide. The effective management and public health control of this disease depends on early detection and careful treatment monitoring. For many years, the microscopy-based analysis of sputum smears has been the most common method to detect and quantify *Mycobacterium tuberculosis* (Mtb) bacteria. Nonetheless, this form of analysis is a challenging procedure since sputum examination can only be reliably performed by trained personnel with rigorous quality control systems in place. Additionally, it is affected by subjective judgement. Furthermore, although fluorescence-based sample staining methods have made the procedure easier in recent years, the microscopic examination of sputum is a time-consuming operation. Over the past two decades, attempts have been made to automate this practice. Most approaches have focused on establishing an automated method of diagnosis, while others have centred on measuring the bacterial load or detecting and localising Mtb cells for further research on the phenotypic characteristics of their morphology. The literature has incorporated machine learning (ML) and computer vision approaches as part of the methodology to achieve these goals. In this review, we first gathered publicly available TB sputum smear microscopy image sets and analysed the disparities in these datasets. Thereafter, we analysed the most common evaluation metrics used to assess the efficacy of each method in its particular field. Finally, we generated comprehensive summaries of prior work on ML and deep learning (DL) methods for automated TB detection, including a review of their limitations.

Keywords: microscopy; machine learning; *Mycobacterium tuberculosis*; automated medical diagnosis; cell detection; fluorescence; brightfield; classification; regression; segmentation

check for updates

Citation: Zachariou, M.; Arandjelović, O.; Sloan, D.J. Automated Methods for Tuberculosis Detection/Diagnosis: A Literature Review. *BioMedInformatics* **2023**, *3*, 724–751. https://doi.org/10.3390/biomedinformatics3030047

Academic Editor: Alexandre G. De Brevern

Received: 17 July 2023
Revised: 5 August 2023
Accepted: 21 August 2023
Published: 1 September 2023

1. Introduction

Globally, tuberculosis (TB) is the leading infectious cause of death worldwide [1]. *Mycobacterium tuberculosis* (Mtb) is the causative bacteria of TB, which is spread by droplets and aerosols dispersed through coughing. Up to 85% of cases affect the lungs, which is called pulmonary TB. Other, extrapulmonary organs or tissues, such as the brain, kidneys, bone, and skin can also be affected by TB. The research presented in this paper has an emphasis on pulmonary TB as it focuses on the automated analysis of digital images generated from sputum smears to detect the Mtb bacteria that cause TB. According to the World Health Organisation (WHO), up to 2 billion people worldwide have Mtb bacteria in their bodies, with up to 10 million instances of active illness and 2 million deaths every year [2]. For several decades, TB has been treatable with antibiotics, but the emergence of drug-resistant bacterial strains, such as multi-drug resistant (MDR) and extensively drug resistant (XDR) TB, is making antibiotic treatment more difficult [3]. The largest burden of morbidity and death from TB occurs in poor and middle-income nations, where healthcare resources are limited [4]. The early detection of TB improves a patient's chances of successful treatment and recovery while also reducing transmission and lowering the risk that drug-resistant pathogens will emerge [2,5,6].

The slower progress that has been achieved in the automation of TB microscopy can be reflective of several factors [7]. The automation of TB microscopy is a difficult problem to tackle due to, for example, the challenges associated with segmenting and tracking Mtb cells due to their irregular shape and tendency to clump together [8]. 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 smear microscopy methods. However, we contend that the 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. Secondly, microscopy still has a role, not yet supplanted by any other method, in treatment monitoring. Thirdly, and perhaps most importantly, microscopy may have a specific research value in understanding heterogeneity in TB treatment response at the level of individual cell morphology [8,9]. There is no replacement technology for this function in the foreseeable future. There are clear examples where detailed microscopy-based studies of single cells have led to important advances in our understanding of crucial questions about pathogens that cause human infection [10]. Multicolour fluorescence microscopy has contributed to the elucidation of the developmental morphologies of the malaria parasite Plasmodium falciparum [10]. High-content confocal microscopy imaging has been successfully performed to help identify the factors that influence disease severity in infections caused by M. abscessus, an organism in the same bacterial family as Mtb [11]. These results illustrate the potential applications for similar tools in TB.

2. Importance of Microscopy

Sputum smear microscopy has traditionally been the primary method for diagnosing TB. In this method, sputum samples from individuals whose presumptive TB symptoms include a productive cough are thinly smeared, heat-fixed onto slides, and stained using special laboratory techniques. The staining procedure depends on the unusual properties of Mtb cells, particularly their very thick, lipid-rich cell wall, which takes up selective dyes and then resists decolourisation with a dilute acid rinse. The result is that the dye becomes concentrated in short, rod-like structures, approximately 0.2–0.5 $\times$ 1.0^{-7} µm in size, called acid-fast bacilli (AFB) [12–14]. Not all AFB are Mtb cells, but, in the correct clinical context, the detection of AFB in sputum is highly suggestive of pulmonary TB.

There are two main staining and microscopy approaches that are used to visualize AFB. The traditional approach, called the Ziehl–Neelsen method, uses carbol fuschin to label AFB red against a blue background. A conventional, brightfield light microscope at $\times$1000 magnification is used to examine the sputum. Newer fluoresence-based protocols use auramine O to stain AFB bright yellow-green against a black background. A fluorescence microscope (using, for example, a light-emitting diode or laser as a light source to excite the auramine O) is required for this approach, but sputum can be examined at $\times$400 magnification, and the overall procedure is faster [13]. Semi-quantitative grading methods have been created to estimate the bacterial burden in a patient's lungs from the concentration of AFB seen in each microscopic field of view. The findings of sputum smear microscopy are often described as 'negative', 'scanty', '1+', '2+', or '3+', with each successive categorization indicating a higher bacterial load [12].

Although sputum smear microscopy has long been the mainstay of pulmonary TB diagnosis, recent advances in molecular microbiology have resulted in many centres across the globe switching their attention away from smear microscopy and toward polymerase chain reaction-based technologies (such as the Xpert MTB/RIF assay) for TB diagnosis in recent years [15]. Molecular methods have several advantages: they are faster, less subjective, provide more precise bacterial identification, and also supply genetic information on the likelihood of antibiotic resistance in the Mtb cells that they find [2,15].

It would be wrong, however, to suggest that smear microscopy is now redundant. Even when heavily subsidized by donors, molecular diagnostics are expensive, and the WHO has not endorsed the Xpert MTB/RIF assay for treatment monitoring because the

test remains positive in some patients even when therapy is working well [2]. Sputum smear microscopy, on the other hand, is lower-cost, and gradings can remain effective for triaging baseline disease severity and prognosis, with possible implications for treatment inividualisation. At present, smear microscopy remains the WHO's recommended tool for monitoring treatment response [2,15].

Mtb are slowly replicating bacteria, so smear microscopy provides data considerably faster than waiting for organisms to grow in culture, which is the conventional gold standard for TB diagnosis in clinical microbiology practice [12]. When done well, microscopy has a high specificity (99%) for detecting Mtb cells, and it has also become more sensitive since switching from Ziehl–Neelsen to fluorescent auramine-based methods (from 0.34–0.94 to 0.52–0.97 according to one systematic review) [16,17]. The wide ranges of diagnostic sensitivity described in that paper also reflect the technique's complexity and subjectivity. Some researchers also use microscopy to measure changes in the appearance (e.g., size, shape, and lipid content) of individual Mtb cells during TB treatment in order to better understand the changes seen in bacterial phenotype during therapy [18].

Disadvantages of Microscopy and Motivation for Computer-Based Automatic Detection

This paper has already touched on some of the obstacles to using microscopy effectively for clinical patient management and microbiological research on Mtb. For microscopists, maintaining a high level of skill necessitates a consistent commitment of time. To stay proficient, WHO guidelines recommend that practitioners should study at least 25 slides every day [2]. Examining a slide is complex: viewed down the microscope, each slide is sub-divided into tiny fields of view (FOV), which must be examined one by one. Human error, fatigue, and subjective decision making are certain to impair specificity and sensitivity results. Some slides are difficult to interpret because some AFB have unusual appearances; additionally, some non-bacterial components (artefacts) inside the sputum matrix resemble Mtb cells and may be misidentified as such. Artificial intelligence techniques may present a means to surmount some of these challenges. This paper aims to explore and answer the following questions regarding tuberculosis artificial intelligence (TB–AI):

- What datasets of TB microscopy images are available online, and what microscopy methods were used to generate them?
- What challenges to the development of AI image analysis methods are presented by the level of variability in currently available TB microscopy image datasets?
- What metrics have been employed previously to assess the efficacy of AI techniques in the analysis of TB microscopy images, and what are their respective advantages and limitations?
- What specific machine learning (ML) and deep learning (DL) techniques have been performed for TB microscopy image analysis, and what knowledge can be distilled from their applications in approaches that did and did not work?

3. Datasets

In order to access and review relevant data on TB–AI, a detailed screening was undertaken of several academic databases. Systematic searches were carried out 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 January 2023. No restrictions were placed on the year of publication. The scientific publications identified from these searches were screened by title, abstract, and full text. Those that were written in English and described the 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.edu, 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 inclu-

sion were also screened to identify any additional datasets that may have been missed by this search strategy. Figure 1 provides an overview of the aforementioned procedure.

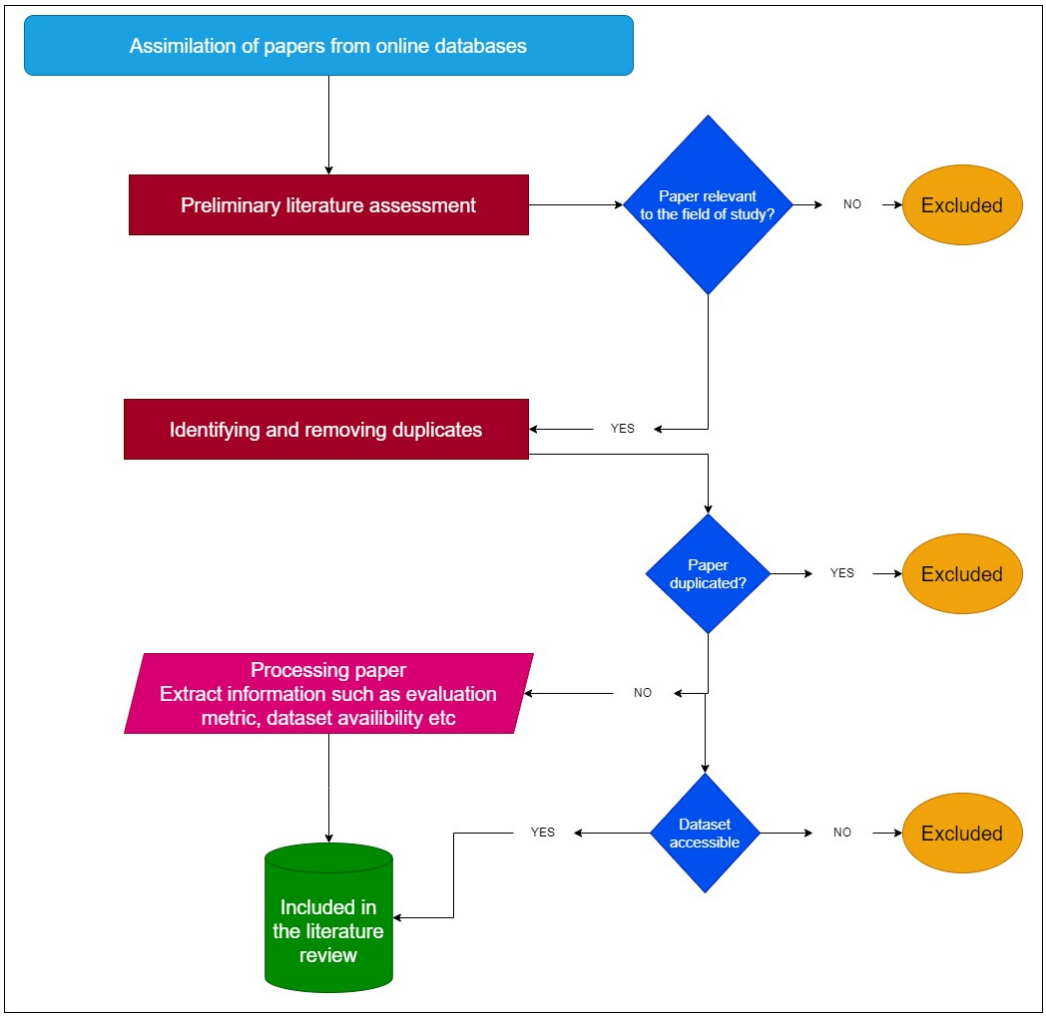


Figure 1. A comprehensive outline of the methodology employed for collating all pertinent information presented in this work.

Once all publications containing relevant datasets were identified for inclusion, the metadata were mapped for each publication, including whether the database used for analysis was currently openly accessible online (see Table 1). The supplementary metadata included the geographical origin of the publication and the specific microscopy technique employed. Furthermore, a record was made of any quantitative evaluation metrics that were stated for each work, along with their corresponding values. This was carried out to facilitate a succinct and cohesive comparison between methods, where applicable.

As described above, conventional brightfield and fluorescence microscopy are used for Mtb bacterium visualisation; consequently, TB–AI research utilises datasets of images derived from both of these methods. Although fluorescence microscopy increases sensitivity, it can also lower specificity since fluorescence makes bacteria and other elements in sputum more apparent; some of these other elements may be mistaken for Mtb cells as a result. Fluorescence microscopy has only become widely available as a TB diagnostic tools in many centres over the last decade and has higher operational costs than brightfield microscopy. A combination of these factors may explain why, particularly in older studies, image datasets are based on brightfield microscopy [19,20].

To the best of our knowledge, there are currently five databases for Mtb sputum smears that are accessible online. These databases are as follows in chronological order: the CDC Public Health Image Library (PHIL) [21], Kaggle Tuberculosis Image Dataset [22], TB_IMAGES_DB_BACILLI.V1 [23], the Ziehl–Neelsen sputum smear microscopy image database (ZSNM-iDB) [24], and lastly, another database collected from brightfield microscopy sputum smears(TBDB) [25]. Only 5 (11%) studies used one of these datasets that are currently available online, while the remainder of the work reviewed in this paper made use of individually owned proprietary datasets. Of course, research based on proprietary datasets is still very 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 2 presents additional details pertaining to the aforementioned datasets.

Table 1. Datasets. Only 20% of all datasets in automated TB research were based on fluorescence-stained sputum smears, suggesting that the vast majority of datasets used brightfield microscopy and were proprietary. In addition, 88% of the fluorescence datasets were used in Europe, and only two were used in the Americas and Asia.

Paper	Year	Microscopy Type	Region of Image Generation	Region of Method Development	Purpose of Research	AI Method Used	Dataset Online
Veropoulos et al. [26]	1998	Fluorescence	N/A	Europe	Diagnosis	ML	No
Forero-Vargas et al. [27]	2002	Brightfield	N/A	Europe	Detection	ML	No
Forero et al. [28]	2003	Fluorescence	Europe	Europe	Detection	ML	No
Forero et al. [29]	2004	Fluorescence	Europe	Europe	Detection	ML	No
Forero et al. [30]	2006	Fluorescence	Europe	Europe	Detection	ML	No
Sadaphal et al. [31]	2008	Brightfield	America	America	Detection	ML	Yes [21]
Costa et al. [32]	2008	Brightfield	America	America	Detection	ML	No
Makkapati et al. [20]	2009	Brightfield	N/A	Asia	Detection	ML	No
Sotaquŕa et al. [33]	2009	Brightfield	America	America	Quantification	DL	No
Khutalang et al. [34]	2010	Brightfield	Africa	Africa	Detection	ML	No
Osman et al. [35]	2010	Brightfield	Asia	Asia	Diagnosis	ML	No
Osman et al. [36]	2010	Brightfield	Asia	Asia	Diagnosis	ML	No
Osman et al. [37]	2010	Brightfield	Asia	Asia	Diagnosis	ML	No
Zhai et al. [38]	2010	Brightfield	N/A	Asia	Detection	ML	No
Nayak et al. [39]	2010	Brightfield	Asia	Asia	Quantification	DL	No
Chang et al. [40]	2012	Flueorescence	Africa	America	Diagnosis	ML	No
Costa-Filho et al. [41]	2012	Brightfield	America	America	Detection	ML	Yes [23]
Santiago-mozos et al. [42]	2014	Brightfield	N/A	Europe	Diagnosis	ML	No
Ayas & Ekinci [43]	2014	Brightfield	Asia	Asia	Detection	ML	No
Costa-Filho et al. [44]	2015	Brightfield	America	America	Detection	ML	Yes [23]
Govindan et al. [45]	2015	Brightfield	America	Asia	Detection	ML	Yes (partially) [21]
Gosh & Nasipuri [46]	2016	Brightfield	Asia	Asia	Diagnosis	ML	No
Priya et al. [47]	2016	Brightfield	Africa	Asia	Detection	ML	No
Soans et al. [48]	2016	Brightfield	N/A	Africa	Quantification	DL	No
López et al. [49]	2017	Brightfield	N/A	America	Detection	DL	No
Yan & Zhuang [50]	2018	Brightfield	Asia	Asia	Detection	ML	Yes [23]
Kant & Srivastava [3]	2018	Brightfield	N/A	Asia	Diagnosis	DL	No
Panicker et al. [51]	2018	Brightfield	America	Asia	Detection	DL	Yes

Table 1. *Cont.*

Paper	Year	Microscopy Type	Region of Image Generation	Region of Method Development	Purpose of Research	AI Method Used	Dataset Online
Samuel & Kanna [52]	2018	Brightfield	Asia	Asia	Detection	DL	Yes
Xiong et al. [53]	2018	Brightfield	Asia	Asia	Diagnosis	DL	No
Mithra & Emmanuel [54]	2018	Brightfield	Asia	Asia	Quantification	DL	Yes [24]
Díaz-Huerta et al. [55]	2019	Brightfield	America	America	Detection	ML	No
Ahmed et al. [56]	2019	Brightfield	N/A	Asia	Diagnosis	DL	No
Hu et al. [57]	2019	Brightfield	Asia	Asia	Diagnosis	DL	No
El-Melegy et al. [19]	2019	Brightfield	Asia	Africa	Detection	DL	No
Mithra & Emmanuel [54]	2019	Brightfield	Asia	Asia	Diagnosis	DL	Yes [24]
Vente et al. [58]	2019	Fluorescence	Africa	Europe	Quantification	DL	No
Yousefi et al. [59]	2020	Brightfield	N/A	America	Detection	ML	No
Serrão et al. [60]	2020	Brightfield	America	America	Detection	DL	No
Swetha et al. [61]	2020	Brightfield	N/A	Asia	Diagnosis	DL	No
Zachariou et al. [62]	2022	Fluorescence	Africa	Europe	Detection	DL	No
Zachariou et al. [63]	2022	Fluorescence	Africa	Europe	Quantification	DL	No

Challenges with Dataset Standardisation

Irrespective of the dataset used, a consistent challenge when analysing TB microscopy image is that the process of sputum smear preparation and image capture can be problematic to standardise. Even when carefully written standard operating procedures are meticulously followed, expectorated sputum is variable in consistency and difficult to homogenise. This affects the thickness of smears and consequently influences the degree of background material and stain uptake on microscopy slides (see Figure 2). Once slides have been prepared, the process of reading them comprises magnification (typically from $\times$400 to $\times$1000) and sequential examination of small FOVs. At this stage, when researchers are preparing collections of FOVs for automated analysis, procedures vary. The most common options are: (i) the manual inspection and creation of an image set by an expert [58,63], (ii) auto-focus algorithms [3,29,38], or (iii) successive cropping of the whole slide followed by a filtering stage to remove FOVs void of bacteria [62]. Individual fields of view, or subsections of them, are often additionally cropped into even smaller images, wherein bacteria are present. All of these methods ultimately utilise FOVs of arbitrary dimensions, i.e., there are 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. Additionally, researchers in different settings may regularly have different hardware (e.g., digital cameras with different specifications).

This all has implications for downstream biological research based on image interpretation. For example, knowing and comparing the physical size of a bacterium under different physiological or treatment conditions may be useful for research into the effectiveness of TB therapy, but measurement of this is impossible if image dimensions and magnification are not standardized and recorded at the time of image collection. Studies using online accessible image sets illustrate this problem. Yan et al. [50] evaluated their approach to Mtb cell detection from Ziehl–Neelsen-stained smears on their own proprietary dataset and the online ZNSM-iDB dataset [24], with the latter yielding much lower accuracy because the dimensions and resolutions vary considerably within the ZNSM-iDB images.

Table 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 its labels are constructed.

Image Dataset Name	URL	Content of Dataset	Image Annotation	Label Type
CDC Public Health Image Library [21]	phil.cdc.gov (accessed on 22 August 2023)	Microscopy images within general collection of TB-related images, 25 brightfield slides 15 fluorescence slides	None	N/A
Kaggle Tuberculosis Image Dataset [22]	kaggle.com/datasets/saife245/tuberculosis-image-datasets (accessed on 22 August 2023)	1265 brightfield images	Yes	Bounding Boxes
TB_IMAGES_DB_BACILLI.V1 [23]	Free access can be applied for at tbimages.ufam.edu.br (accessed on 22 August 2023)	120 brighfield images	Yes	Bounding Boxes
ZNSM-iDB [24]	drive.google.com/drive/folders/1HPcJzwKi76WwCFYj7dHUgVA31dAyFyTF (accessed on 22 August 2023)	9 sets of brightfield images (50–90 images per set)	Yes	Bounding Boxes
TBDB [25]	Freely available by contacting the authors	3102 brightfield images	Yes	Not specified

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

Figure 2. The background/bacteria contrast and magnification levels of two fluorescence images prepared with Nile red staining are different between the two images. For (**a**), the magnification is ×1000, while for (**b**), it is ×800. White arrows indicate the presence of bacteria

ML and DL are data-driven, and the majority of methodologies employed suffer from a lack of an appropriate volume of training data, which has a substantial impact on performance. The fact that so few openly available TB smear microscopy datasets exist and the absence of standardisation in the methods used to generate them may have contributed to a situation where the majority of publications in the TB–AI field utilise their own dataset both for training and the evaluation of their methods. This reduces the likelihood that the results of one method can be easily replicated within other settings. Approaches that generate promising results in one dataset may not do so on another. Therefore, consideration should be given to whether it is possible to establish databases of microscopical images according to agreed standardised protocols and parameters. Although this would be desirable, it may be difficult to achieve because some of the causes of variability between datasets outlined above are hard to eliminate.

4. Evaluation of Performance Metrics

In the development of any new method, the formulation of evaluation metrics that reflect the industry standard is required so that the performance of a novel method can be compared to the current state of the art. In most of the research on TB detection, classification is used to assess a method's ability to discriminate between FOVs that contain Mtb bacteria and those that do not. Other works choose to employ a segmentation step prior to classification, during which they locate Mtb bacteria. The primary distinction between the two tasks is that the former involves making a diagnostic decision based on an FOV, whereas the latter determines whether or not individual objects belong to the bacterial class. Instead of making a binary decision on whether a certain FOV includes Mtb bacteria or not, some methods also take a quantitative approach, using regression to count the number of bacteria present in each FOV. We sought to compile a set of performance measurements for each category of applicable techniques, including classification, segmentation, and regression. These performance metrics convey empirical information when applied to a model's test set, but they are also sometimes applied to a model's training data to draw further inferences about the model's behaviour.

4.1. Classification Metrics

The most commonly utilised evaluation metric for any method of making a binary classification for a given medical diagnosis is its capacity to accurately differentiate between positive and negative occurrences [64]. This is also true in TB detection/diagnosis research. It is noteworthy that the terminology associated with the two aforementioned words (diagnosis and detection) differs in context. Diagnosis refers to the ability of a given test or

method to distinguish between samples from potential patients, indicating whether they are positive for the disease (i.e., afflicted) or negative for the disease (i.e., not afflicted). Detection refers to the capacity of a given method to precisely determine the location of Mtb bacteria or any other type of bacteria, for that matter. Consequently, it is possible for an automated diagnostic approach to yield an accurate outcome (positive or negative) for all the erroneous rationales, wherein the verdict is not grounded on the detection of bacteria but rather on other factors. Consequently, this method cannot be considered a tool for detection. In addition to accuracy, which consists of correct vs. incorrect outcomes, it is important to describe four groups of possible outcomes. The successfully categorised groups are the accurately predicted true-positive (TP) and true-negative (TN) model outcomes, where, i.e., instances of the two classes are correctly classified. In contrast, instances of the negative class that were incorrectly predicted as positive are known as false positives (FP). Similarly, false negatives (FN) relate to positive class members who were incorrectly predicted as negatives. The definition of accuracy is:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

However, accuracy is often insufficient for evaluating medical diagnostic models. Its shortcomings are obvious when assessing models with imbalanced datasets, which is the situation for the majority of medical artificial intelligence (AI) applications of data sampling and the properties of the domain. It is possible that imbalances in the instances across the classes are caused by the way they were collected or sampled from the problem domain. Sensitivity (also known as recall) and specificity are also very significant metrics that are frequently employed in TB diagnostics/detection research. Sensitivity is used to assess model performance since it reveals the number of positive occurrences that the model accurately predicted. A model with high sensitivity will have few false negatives, meaning it will miss some positive examples. Sensitivity is defined as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

Consider the case of a disease medical test. Specificity refers to a test's capacity to exclude individuals without a disease. The specificity of a test is the fraction of individuals who test negative for the ailment who actually do not have it. This statement may also be written as:

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

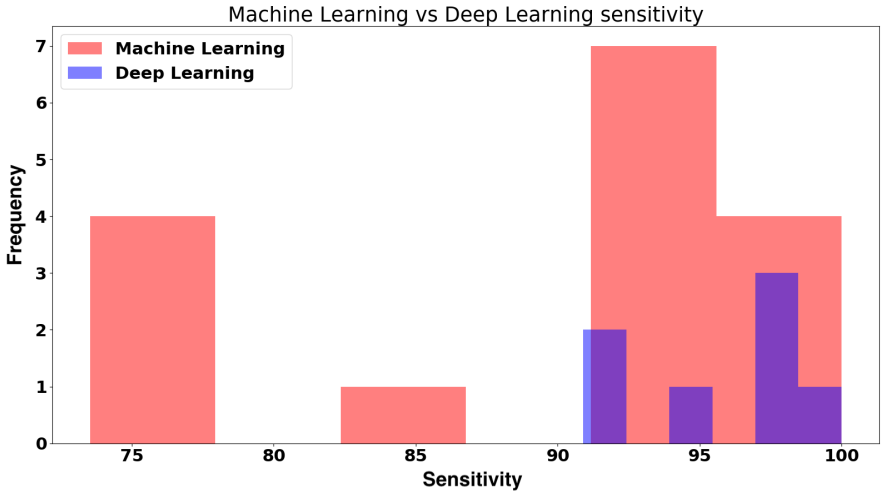
Despite the fact that there are fewer studies that utilise DL than ML in TB, they have been consistently scoring higher in sensitivity and specificity, as shown in Figure 3.

Precision measures the proportion of the positively predicted occurrences that were accurately classified. Precision is defined as follows:

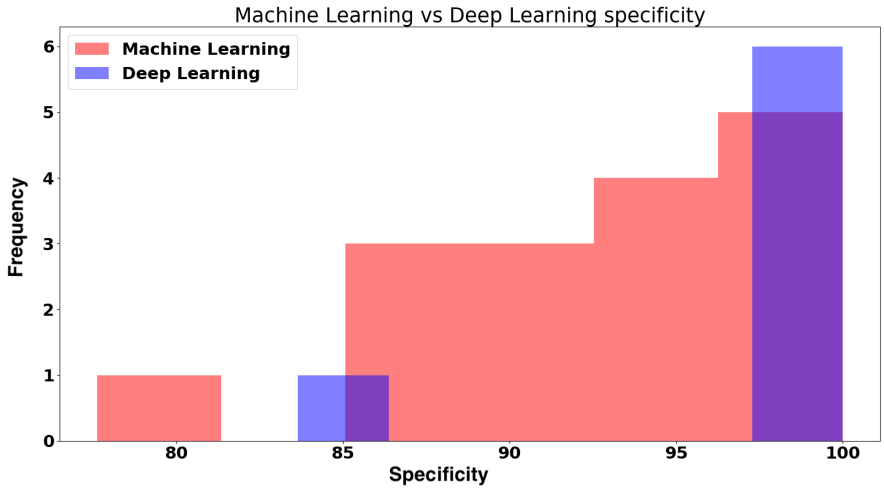
$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Precision is also known as the predictive positive rate (PPR), which is often accompanied by receiver operating characteristics (ROC) and the area under the ROC curve (AUC). When the context of the task focuses more on accurately identifying positive samples and less on accurately identifying negative samples, it is important to provide an informative analysis using the PPR metric [49,52,62]. These measures assist in the analysis of the trade-off between the true-positive rate (TPR), commonly referred to as sensitivity, and the false-positive rate (FPR), also known as the complement of specificity, across the different decision thresholds of a binary classification model. The FPR is defined as:

$$FPR = \frac{FP}{FP + TN} \tag{5}$$



(**a**)



(**b**)

Figure 3. Comparative analysis of the sensitivity (**a**) and specificity (**b**) attained by works utilising ML and DL. Even though they are fewer in number, DL algorithms routinely score > 90 on both sensitivity and specificity.

The AUC is a scalar value that represents the area under the ROC curve. It summarizes the overall performance of the model across all classification thresholds. The AUC value ranges from 0 to 1, where:

AUC $= 0.5$: The model fails to exhibit superior performance when compared to random guesses.

AUC > 0.5 : The model 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) enables the combination of precision and recall into a single metric that encompasses both characteristics. Neither precision nor recall alone provide a comprehensive explanation of a model's performance. Indeed, a model may reach perfect (or near-perfect) precision while its recall remains low, and vice versa. The F-measure enables the merging of the two aforementioned metrics into one score. After

calculating precision and recall, the two scores can be combined to determine the F-measure, which is defined as:

$$F - measure = \frac{(2 \cdot Precision \cdot Recall)}{Precision + Recall} \tag{6}$$

4.2. Regression Metrics

Some research concentrates on estimating the overall number of bacteria present within a given FOV with the end objective of quantifying bacterial load as a contributor to illness severity in patients. Multiple works have chosen regression analysis for the aforementioned tasks, therefore forecasting a real number even when the actual count can only be an integer [58,63]. This choice is prompted by the desire to preserve information regarding the uncertainty involved in deducing the bacterial count. Typically, the mean average error (MAE), mean squared error (MSE), and coefficient of determination (R^2) are employed. Taking the average of all observations, the MAE measures the absolute distance between the observations (the images of the dataset) and the regression predictions. The absolute value of these distances is used to correctly account for negative errors. MAE is mathematically expressed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\mathrm{y}_i^{true} - \mathrm{y}_i^{pred}\right| \tag{7}$$

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. Moreover, the MAPE demonstrates robustness in instances where there are outliers or extreme values within the dataset. This is due to its emphasis on the relative errors rather than the absolute magnitude of errors. This particular attribute renders it a valuable measure for practical contexts in which deviations are prevalent and the precise prediction of such exceptional values is of utmost importance. The MAPE is defined as:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\mathrm{y}_i^{true} - \mathrm{y}_i^{pred}}{\mathrm{y}_i^{true}}\right| \times 100 \tag{8}$$

In contrast, the distance may be squared to provide differentiability in all instances of outcomes; this process makes it easier to perform mathematical operations in comparison to a non-differentiable function, such as the MAE. One of the major drawbacks of the MAE is that it cannot be differentiated at zero. Numerous optimization techniques often employ differentiation to obtain the optimal parameter values for the evaluation metric. It may be difficult to calculate gradients in the MAE. Absolute distances are removed and each distance is squared to define the MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{y}_i^{true} - \mathrm{y}_i^{pred}\right)^2 \tag{9}$$

The key difference between the MSE and MAE is how each penalises errors caused by comparing the predicted data to the ground-truth data. Since the MSE is a squared error, it penalises large errors more heavily than absolute error since the errors are squared rather than just calculated as a difference. Therefore, the MAE is not sensitive to the outliers

within a given dataset. Consequently, the robustness of each metric and when it should be used is contingent on the nature of the task being evaluated.

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{y}_i^{true} - \mathrm{y}_i^{pred})^2} \tag{10}$$

Lastly, represents the fraction of the variance in the dependent variable that the linear regression model explains. It is a scale-free score; therefore, whether the numbers are small or large, R^2 will always be less than one. Therefore, it indicates the predictor variables' ability to explain the variation in the response variable. It can be expressed as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} \tag{11}$$

It is used to describe the extent to which the independent variables in a linear regression model explain the variability of the dependent variable. The value of R^2 always increases as independent variables are added, which may lead to the inclusion of redundant variables in the regression model.

4.3. Segmentation Metrics

Similar to the context of classification, pixel-wise accuracy is a popular criterion in segmentation; however, in this context of image-to-image pixel similarity overlap, it is less relevant and helpful. As the bulk of a microscopy slide or FOV is the background, it is evident that a model may achieve high accuracy without identifying any Mtb bacilli. In actuality, the model is learning to recognise background data rather than objects of interest (bacteria). A more suitable metric is to quantify the overlap of pixel similarity between the predicted segmented images and the ground-truth segmented images. A frequent metric for this specific task is the Sorensen–Dice (SD) coefficient (also known as F1-score), which is defined as follows:

$$SD = \frac{2 \cdot |S_1 \cap S_2|}{|S_1 + S_2|} \tag{12}$$

where S_1 is the number of elements in set 1, i.e., the pixel values in an image. Likewise, S_2 holds the numbers of elements in set 2. Another common metric that is similar to SD is the Jaccard index (also known as the intersection over union (IoU)). As in the concept of (SD), this method computes the degree of pixel similarity overlap between two images, specifically between predicted images and corresponding ground-truth images. The definition of the Jaccard index between two sets is:

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{13}$$

Thus, similar to the SD, the greater the number reached (which can range from 0 to 1), the better. While the two metrics have some similarities, they are different and serve different purposes in image segmentation tasks. Given a value for the SD coefficient S, it is possible to determine the corresponding Jaccard index value J, and vice versa. In general, the Jaccard metric tends to punish single occurrences of incorrect classification more than the SD, even when both metrics agree that a single case is incorrect. The Jaccard index quantifies the degree of similarity between two images by computing the ratio of the size of their intersection to the size of their union. Conversely, the SD coefficient quantifies the proportion of the shared elements (pixel values) between two sets in relation to the

combined total of the elements in both sets. To summarise, the concepts of intersection size and intersection area are interrelated and utilised to quantify distinct aspects of the overlap of regions of interest (ROIs) within an image. The size of the intersection denotes the quantity of shared elements between ROIs given two images, whereas the area of the intersection represents the spatial coverage of that overlap. Therefore, the SD tends to evaluate performance closer to the average, while the Jaccard score measures performance closer to the worst-case scenario. However, from the perspective of averaging these scores across a large number of inferences, both suffer from an additional disadvantage: they both overemphasise the significance of sets with few or no actual ground-truth positive sets. In a typical example of image segmentation, if an image contains just a single pixel of a detectable class and the classifier recognises that pixel and one other pixel, the SD is $2/3$, and the Jaccard index is much worse, at $1/2$. Such errors may significantly affect the average score for a series of images. In conclusion, each pixel inaccuracy is weighted inversely proportionately to the size of the ROIs between two images as opposed to being treated equally.

The evaluation of the proximity of a structure's perimeter is an alternative method. Given model-predicted images with highlighted objects of interest and their matching ground-truth images, the closer the distance between these structures, the more similar they are. The Hausdorff distance quantifies the distance between two subsets of a metric space, thus transforming a set of non-empty compact subsets of a metric space into its own metric space. The Hausdorff distance between two point sets, S_1 and S_2, is defined as:

$$H(S_1, S_2) = max(h(S_1, S_2), h(S_2, S_1)) \tag{14}$$

where $h(S_1, S_2)$ is defined as:

$$h(S_1, S_2) = \underset{s_1 \in S_1 s_2 \in S_2}{max\ min} \|s_1 - s_2\| \tag{15}$$

and 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 as the norm. In certain instances, the traditional Hausdorff distance may result in unfair performance assessments due to the fact that it penalises single outliers. Additionally, there are some studies that have used the Hausdorff distance as a bidirectional metric for image comparison and evaluation, including in the context of the modified William index (MWI) [65]. 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. In fact, the authors used the Hausdorff distance in their work to determine the distance between each predicted structure and the actual structure in a given set of images [66]. Table 3 provides insight into what evaluation metrics are used by works that include an evaluated segmentation stage.

Table 3. 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	Hausdorff Distance	Jaccard Index	SD
Khutlang et al. [66]	0.96	N/A	N/A
Soans et al. [48]	0.06	N/A	87%
Diaz-Huerta et al. [55]	N/A	96%	N/A
Mithra & Sam Emmanuel [54]	N/A	95%	N/A
Zachariou et al. [63]	N/A	94%	89%

5. Research Utilising ML

In this section, we will provide a synopsis of all the research that has been conducted employing ML algorithms in conjunction with computer vision techniques. Anything that does not adhere to the convention of deep convolutional neural networks (DCNN) is regarded as traditional ML; it is therefore included in this section. Table 4 shows a summary of the most common evaluation metrics used by all papers included in this section.

Table 4. This table summarises the outcomes of the most frequently employed evaluation metrics, namely accuracy, sensitivity, and specificity. There is considerable variation between the evaluation metrics used between different studies, emphasising that there is no accepted gold standard and illustrating the difficulty of comparing research carried out in different settings.

Paper	Accuracy	Sensitivity/Recall	Specificity
Veropoulos et al. [26]	97.90%	94.10%	99.10%
Forero-Vargas et al. [27]	N/A	N/A	91.00%
Forero et al. [28]	N/A	93.30%	91.68%
Forero et al. [29]	N/A	86.66%	99.74%
Forero et al. [30]	N/A	94.67%	98.10%
Sadaphal et al. [31]	N/A	N/A	N/A
Costa et al. [32]	N/A	76.65%	88.65%
Makkapati et al. [20]	N/A	N/A	N/A
Khutalang et al. [66]	86.85%	99.95%	77.62%
Osman et al. [36]	86.32%	N/A	N/A
Osman et al. [35]	98.07%	100.00%	96.19%
Osman et al. [37]	N/A	N/A	N/A
Zhai et al. [38]	N/A	100.00%	94.00%
Chang et al. [40]	N/A	92.30%	88.00%
Santiago-Mozos et al. [42]	N/A	73.53%	99.99%
Ayas et al. [43]	N/A	75.77%	96.97%
Costa-Filho et al. [16]	91.45%	93.41%	89.50%
Costa-Filho et al. [44]	93.25%	93.75%	88.46%
Govindan et al. [45]	N/A	72.89%	N/A
Gosh et al. [46]	N/A	93.90%	88.20%
Priya et al. [47]	91.30%	91.59%	88.46%
Aymas et al. [67]	70.52%	N/A	N/A
Yan et al. [50]	N/A	97.46%	93.99%
Diaz-Huerta et al. [55]	98.66%	N/A	N/A

5.1. Image Gradient-Based Approaches

The principle underpinning the image gradient-based approach 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 latter method uses graphical displays of data, such as histograms, to analyse pixel hue bands in order to establish what colour range bacteria are more usually found in. This is often the initial stage of the procedure since the eventual aim is to convert images into binary masks, fully erasing the background and with likely microorganisms as white contours. In the next stage, a shape descriptor is employed to extract the characteristics of the shape of the bacilli. The approach thereafter differs, with some research opting to utilise a classifier, while other

studies attempted to manually deduce the difficult requirements with the use of heuristic information on bacterial shape.

Veropoulos et al. work [26] was possibly the first published paper towards automatic TB diagnosis. He devised a five-step methodology, combining computer vision techniques with a simple neural network as a classifier. First, a Canny edge detector was employed to detect object boundaries and save image processing time. Pixel linking was used to fix damaged structures caused by noise, and then the resulting image was translated from its spatial domain to its frequency domain using discrete Fourier transform (DFT). Following the calculation of Fourier coefficients to serve as form descriptors for bacteria, these coefficients were input into four kinds of classifiers: K-nearest neighbours, a neural network, a Kernel–Adatron algorithm [68], and a support vector machine (SVM). The best performance measure recorded by the authors was 97.9% accuracy. Even though this work considered bacteria in their simplest form (i.e., a singular elongated structure), its most significant contribution is that it indicates the feasibility of TB detection using computer-aided image analysis. In fact, another study published more than a decade later used the Canny edge detector as its primary approach for identifying bacteria [42]. In addition, an extra pre-processing step with an adjustable colour threshold was established for the green colour component of the image. Then, two successive SVM classifiers were used, the first of which discarded incorrectly identified objects from the previous stage and the second of which classified these objects based on their pixel value. The former classifier employed a collection of rotation and translation invariant characteristics of each candidate object as input.

Forero et al. [28] used a similar method, which included a segmentation phase comprised of a Canny edge detector, morphological operators, and the classification of the resulting image. Different bacilli characterisation and the use of just clustering approaches for the classification part are two major variations. Forero et al. [29] published yet another work using a similar concept, with the significant distinction being that autofocus algorithms were utilised for the magnification levels and construction of FOVs. Despite comparable or somewhat worse results than the two preceding attempts [26,28], this was the first study to implement automated FOV generation. Next, Forero et al. [30] released a further work that categorised colour thresholding and form descriptors using clustering algorithms. However, they used Gaussian mixture models this time since they were able to create a distribution of class features. Therefore, an instance of the bacterium class is represented in the feature space as a mixture of Gaussians. Although the results were comparable with their previous results, the dataset was produced manually using a fluorescence microscope.

In addition, a further study using colour thresholds in the segmentation stage was published, in which the authors were able to isolate bacteria from an image's background [32]. They subtracted the red and green channels from an RGB image and determined a threshold value that distinguished objects of interest from the background. The absence of a classification step prompted the authors to develop a heuristic filtering stage. Another method used colour thresholding as their principal technique in bacterium segmentation [20]. Their proposed approach was to select the hue range x°–360°, where x is an adaptable number dependent on the input image. Similarly, no classification technique was used, only a filtering stage utilising heuristic knowledge of bacterium morphology characterisation. The authors reported no evaluation metrics. In both works, it is evident that the absence of an automated classifier had a harmful impact on the results in comparison to earlier works [26,28–30]. Two approaches for segmenting TB images using chromatic information are shown in a third work that does not include a classifier [27]. The first technique is based on the information contained in each distinct chromatic histogram and the fuzzy segmentation of colour images. The second technique is 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 his colleagues published similar works [35,37]. In their first paper, the authors designed a colour-filtering stage utilising the hue range in conjunction with luminance and adaptive parameters [36]. This work was conducted only for the purpose of segmentation and used the k-means clustering technique for testing. Although no evaluation metrics were provided by the authors, the findings indicate that some background still surrounds the bacteria in the images. Using their segmentation technique to expand on their work from the prior paper, where $I = 2$ [36], the resulting segmented image was clustered into background and non-background regions [35]. After calculating the moments of the second and third order, a set of seven Hu invariant moments was generated. The generated features were then fed into a genetic algorithm neural network (GA-NN) for classification. The authors did not report specificity or sensitivity, just 88.54% accuracy in correctly classifying bacteria. In their third article, they used the same segmentation method as in the first, but this time, they employed the geometrical characteristics of Zernike moments [37]. Additionally, a hybrid multi-layered perceptron (HMLP) was used for their final classification stage, which is similar to today's popular Resnet [69] in that it skips connections and adds an identity to the layer activation function. Another work employed colour space transformations in two independent colour spaces, HSV and CIEL * a * b * c, isolating the H and L components from each. An adaptive threshold was used on both of these derived components to distinguish the bacteria from the background.

Remaining within the scope of colour space transformation, the initial approach taken was to create a scalar selection from the following colour spaces: RGB, HSI, YCbCr, and Lab [16]. The components and removal of the components of these colour spaces were employed for pixel classification in the segmentation step. In the second step, a feedforward neural network pixel classifier with selected features as inputs was used to separate bacilli pixels from the 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, the authors released a second paper with the addition of three filters that used RGB space components: rule-based, geometric, and size filters [44]. This combination was then utilised as an input for an SVM and NN. In this work, the authors improved their sensitivity results from 91.5% to 96.80%. Yan et al. retrieved channel a from the Lab space and then extracted the edges (bacterial structures) using a gradient threshold [50]. In addition, the aspect ratio, circularity, and area were employed to eliminate incorrectly detected structures.

Using just the RGB space, the authors defined conditions on each different component of the space that best met the criteria for distinguishing bacteria from the background in a binary image [46]. To eliminate false contours, i.e., predicted 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 belonged 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 from the image [47]. After the image was segmented, the border shape of the areas of interest was characterised by 15 Fourier descriptors (FDs), with the most prominent being chosen using fuzzy entropy measures. These particular FDs of the TB objects were input to the SVM learning algorithm of an MLP neural network.

Yousefi et al. [59] suggested a novel statistical model of the form and colour of TB bacilli in Ziehl–Neelsen-stained light microscope images in order to detect the bacilli in these images. These basic statistical models were used as a universal library for rebuilding any bacillus with different background colours and may overcome the challenges associated with geometric feature extraction techniques. Based on the eigenvalues of the shape and colour models, the authors classified the individual bacilli and overlapping bacilli in the rest of the picture using several approaches. The k-NN classifier performed the best among the evaluated classifiers, with an average accuracy of 82.7% for single-bacilli and overlapping

bacilli recognition. In addition, the accuracy of their method for recognising bacteria and overlapping bacteria from artefacts and background was 99.1%. However, based on their imbalanced dataset and results, it seems that their algorithm was trained to classify background rather than bacteria.

5.2. Stochastic-Based Approaches

This section focuses on publications that aimed to develop probabilistic inferences over a given distribution using some type of stochastic-based methodology. In the literature, both unsupervised (such as k-means) and supervised (such as Bayesian classifier) methods have been used. Govindan et al. provided an example of unsupervised learning-based segmentation in which they utilised k-means clustering in conjunction with decorrelation stretching to identify areas of interest [45]. Consequently, dilating and eroding morphological operators were required to close any broken edges in the final segmented image. Fourier descriptors, eccentricity, and compactness were the feature types utilised for contour information extraction. Finally, the candidate contours were classified using an SVM model. Alternatively, a random forest approach, which is a supervised learning method to classify each pixel as a possible bacilli area based on local colour distributions, could be employed [43]. Each pixel was thus labelled as either a prospective TB bacilli pixel or not. Then, each pixel group was rotated, scaled, and centred inside a bounding box before being classified using the described RF learning algorithm trained on manually designated TB bacterium patches in the training images.

Sadaphal et al. serves as a prime example, as it employed Bayesian segmentation based on the a priori knowledge of bacterial colour [31]. In addition, after the application of morphological operations, a set of shape criteria evaluated whether predicted objects of interest belonged to the bacteria class, were probable bacteria, or were not bacteria. These criteria included the ratio of axis length, eccentricity, and area. A similar method was described in which two Bayesian pixel classifiers were used to partition brightfield microscopy images into background and bacteria [66]. The extraction of geometrically transformed invariant features and the optimization of the feature set 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, using a subset of the collected features. The accuracy, sensitivity, and specificity were all reported to be more than 95%. In the same year, they published a second work with a similar two-step approach, but this time, the segmentation was accomplished using a mixture of Gaussian classifiers [34]. As seen by their findings, this method worked best for both stages (segmentation and classification). In this work, the overall sensitivity was increased by over 2%, while both accuracy and specificity were reduced by more than 4%.

Gaussian-based techniques can also be seen in the literature. The authors of one paper in this realm used a white top hat transform and template matching with a Gaussian kernel to binarize images into a black background and white regions of interest. As is usually done, diluting and eroding morphological operators were utilised to close fractured contours. The binarized image was then used for feature extraction using Hu's moments, geometric and photometric features, and histograms of oriented gradients (HOG). Finally, these features were used to classify whether each candidate contour belonged to the bacterium class using an SVM. Another paper using a Bayesian classifier based on a Gaussian mixture model was published [55]. Despite not being promoted as a diagnosis, the authors' work consisted of segmentation, correctly distinguishing bacteria from the background. The last stochastic paper released included adaptive signal-processing approaches, such as the least mean squares and reduced rank with eigendecomposition algorithms, both of which contain learning parameters for optimization during training [67]. Similar to Diaz-Huerta's work, there was no classification *per se* since this was a study focused 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. However, one may argue that the reasons they reduced their dataset are the same as those that inspired the

automated detection of tuberculosis and the difficulties of manually examining TB smear slides. In Section 2, we have already outlined that reading sputum smear microscopy sides is prone to subjective evaluation due to the variability in staining and image generation methods. Therefore, using only high-quality, noise-free images for TB–AI assessment may introduce selection bias to the research and limit comparison with other real-world works that have used a more representative range of images [3,28,30,39,58,63].

6. Research Utilising DL

In this section, all publications whose primary focus was DL, which often includes the use of convolutional neural networks (CNNs) or deep convolutional neural networks (DCNNs), are presented. Using CNNs for feature extraction and/or classification places the task into the gradient-based domain. To deduce the behaviour of a local image in response to sudden shifts in pixel values, individual kernels or receptive units perform sequential locus convolutions on the image. Typically, the detection of Mtb bacteria or TB diagnosis is approached in two ways in the literature. The first typically involves a two-fold approach that utilises an image processing technique, such as Canny edge detection, to pre-process the image. This may include operations such as the binarisation of the image, the contour extraction of objects, or noise removal, as we have seen before in Section 5.1. Rather than utilising generic DCNN models, the alternative approach involves designing a DCNN architecture that is customised to suit the particular task being addressed. The two aforementioned approaches are not inherently mutually exclusive, as proposed methods frequently incorporate a combination of both approaches. Table 5 presents a summary of the predominant evaluation metrics employed by previous methods discussed in this section.

Table 5. Table summarises most common evaluation metrics for DL paper results. In general, works applying DL methods are more recent than those employing ML techniques. As may be deduced, the usage of evaluation metrics has become more standardised in recent years.

Paper	Accuracy	Sensitivity/Recall	Specificity
Lopez et al. [49]	N/A	N/A	N/A
Kant et al. [3]	99.80%	83.78%	N/A
Panicker et al. [51]	N/A	97.13%	N/A
Samuel et al. [52]	95.05%	N/A	N/A
Xiong et al. [53]	N/A	97.94%	83.65%
Ahmed et al. [56]	96.07%	N/A	N/A
Hu et al. [57]	98.40%	98.00%	98.4%
El-Melegy et al. [19]	N/A	98.4%	N/A
Mithra et al. [70]	97.55%	97.86%	98.23%
Serao et al. [60]	99.67%	99.98%	99.34%
Zachariou et al. [62]	N/A	89.02%	100%

6.1. Custom-Made CNN Architectures

Lopez et al. provided a technique for the automated classification of brightfield smear microscopy patches employing RGB, R-G, and greyscale patch versions as inputs to a CNN [49]. A disadvantage of this method is the lack of a detection stage, since the input consisted of small patches containing bacteria (or not). The majority of techniques incorporated a detection phase that enabled the localisation of bacteria, thereby automating a significant portion of an otherwise laborious manual process. This technique does not completely automate the nature of the aforementioned process. Another method involving the training of manually clipped patches using whole slide images (WSI) was published [53]. The training was carried out with a pretrained CNN on the CIFAR-10 dataset, with the

input size set at 32 × 32 pixels. To improve the results, Boostrap 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 cut positive (bacteria containing) and negative patches (void of bacteria) was described by Serrao et al. [60]. Each patch binarisation involved the segregation of background and foreground regions, where the latter corresponded to Mtb bacteria. Finally, the authors combined 100 of these patches into a 400 × 400 pixel mosaic image. These mosaic images were inputs for the three CNNs proposed by the authors. All of the methods listed in this section so far have a key restriction in that they are not fully automated, as they focus on manually cropped, two-digit-sized patches. Zachariou et al., on the other hand, proposed a method to automatically and sequentially slide over the slide to crop FOVs, excluding negative FOVs while keeping all positive ones [62]. They proposed two distinct encoders, one of which was trained to differentiate between bacteria by inferring pixel intensity, while the other was trained to do so by deducing bacterial shape. This approach effectively generated feature maps based on these two criteria. The first feature extraction process involved utilising a greyscale FOV as input. In contrast, the subsequent feature extraction process involved utilising a binarised FOV as input. This was due to the fact that the pixel intensity was no longer relevant in the latter case, as the FOV was now comprised solely of black (background) and white (foreground) elements. Finally, the feature maps of the two encoders were concatenated and input to a third CNN, whose linear layer produces a positive or negative classification result given a FOV.

6.2. Automatic Creation of FOVs

Kant and Srivastava also used a patchwise classifier that categorised whether a particular patch included bacteria or not [3]. In this instance, though, an autofocus method was utilised to construct 20 × 20 pixel patches from the full slide. The CNN used to classify these patches was composed of five convolutional layers and no linear layers. Similarly, Samuel and Kanna presented another paper describing an automated technique for acquiring FOVs from microscope slides [52]. Then, these FOVs were utilised to train a customised InceptionV3 model with transfer learning to derive bacteria inference feature maps. Finally, these feature maps were employed as training data for an SVM to determine whether or not a specific image FOV included bacteria. Next, a method providing a classification approach for complete slides was presented [57]. Taking into account the settings of high-resolution slide (which are usually gigabytes in size), the authors developed a dataset creation technique based on non-overlapping subgraph partition. ResNet [71], InceptionV3 [72], and DenseNet [73] were utilised with transfer learning to assess their method. InceptionV3 fared the best, with a WHO 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 an incorrect count of bacteria, which could cause difficulties when estimating bacterial load.

6.3. Gradient-Based Approaches

In this section, we examine publications that applied a multi-step process in which, in the first transitional stage, the input images were segmented using a gradient-based CV algorithm. Panicker et al., for instance, utilised the fast nonlocal means method to denoise their images, followed by Otsu's threshold to binarize the images into background and foreground [51]. 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 non-standard Mtb shapes, i.e., anything other than elongated rods. In another work [70], the channel area thresholding (CAT) technique was proposed for bacterial image segmentation. The intensity-based local bacilli characteristics were derived utilising a location-oriented histogram and sped-up robust feature (SURF) algorithm extraction. Deep belief neural networks were used to classify the

bacilli items precisely following segmentation. In another similar paper, brightfield sputum images were preprocessed by employing noise reduction and intensity modulation [61]. Their segmentation method solely used CAT in addition to the features, such as the HOG and the SURF, that were extracted. Classification was conducted using a CNN classifier, which classified the bacillus as mild, moderate, or severe depending on the number of pixels classified as belonging to the bacteria class. Although the authors indicated significant sensitivity and specificity, they provided no more information on the architecture of the employed model.

6.4. Employing Existing Models for Mtb Feature Extraction

For this section, we examine papers that employed existing architecture models as the foundation of their methodology, with or without transfer learning. For instance, Ahmed et al. proposed a method in which they categorised numerous bacteria associated with a variety of diseases [56]. To do this, 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. Following this, El-Melgey et al. presented a work in which they utilised a faster region-based CNN (RCNN) to swiftly localise bacteria using ground-truth bounding boxes [19]. However, due to the high likelihood of false positives, the authors introduced a second step to determine whether the projected bounded boxes actually belonged to the bacterium class. The authors presented comparative evaluation results for their methodology; however, it is worth noting that the bounding boxes utilised in this approach were limited in size and could only accommodate a single bacterium. This represented a notable limitation of the method, as bacterial cells frequently became clogged and overlapped with one another.

7. Research on Mtb Bacteria Quantification

As described in Section 4.2, some TB–AI work has been conducted with the aim of not only detecting the presence or absence of Mtb but quantifying the bacterial load within microscopy images. For clinicians, the quantification of the sputum bacterial load in patients with pulmonary TB can help assess disease severity and the likelihood of treatment success [74]. Monitoring changes in bacterial load over time can also help to track treatment success, so there are several potential advantages of this approach. Similar to the preceding sections, some researchers have chosen to utilise a multi-stage approach to quantify bacteria, while others manually segmented and counted the bacteria present. The previously mentioned setting provides a proficient illustration of the approach employed by Sotaquirá et al., whereby sputum smear images were converted into YCbCr and Lab colour spaces and subsequently evaluated for their relative differences [33]. The quantification of bacterial population was ascertained through the computation of the mean size of bacilli, taking into account the image resolution and the pixel count of the segmented image. Aside from their qualitative and visual results, the authors provided no evaluation metrics for any stage. Moreover, the heuristic information at the core of their method was dependent on image dimensions and, as explained in Section 3, it is not prudent to assume that all datasets will include images with the same dimensions. Finally, the limitations of this study are evident, as the manual enumeration of bacteria undermines the objective of streamlining the TB diagnosis and detection process.

Subsequently, Nayak et al. proposed a technique that employed colour segmentation and colour space transformation [39]. They described their approach as a five-step process. These stages are as follows: (i) colour-based segmentation, (ii) connected component labelling, (iii) size thresholding on the resulting contours, and (iv) proximity grouping, followed by (v) size constraints. The contours produced by the aforementioned process were utilised to determine how many bacteria were present. Keeping with this section's aim in mind, the image was segmented, and detected objects of interest were manually counted using the HSI colour model [48]. Given this image alteration, a knowledge database was constructed and passed to a decision tree classifier in order to determine

which HSI component values corresponded to the bacterium class. Lastly, similar to the previous research, proximity groupings and size constraints were used to eliminate false negatives, instances in which the background was incorrectly identified as belonging to the bacterium class. By thresholding the hue range, a hue-colour-component-based approach was utilised to segment bacilli, and morphological characterisation was employed to determine whether or not the bacilli were valid [75]. By thresholding the area, perimeter, and contour characterizations, other artefacts were eliminated. Using the area, perimeter, and shape characteristics, clumps of bacilli were detected. Counting occurred following the segmentation of bacilli and bacilli clusters.

The proposed method was comprised of three steps: segmentation, feature extraction, and classification [54]. The input sputum smear microscopy image was first subjected to a colour space transformation, followed by thresholding to generate a segmented image. The image's length, density, area, and histogram characteristics were collected for FHDT-based classification, which classified contours as low bacilli, non-bacilli, and overlapping bacilli. A function (in this case, a decision tree) of entropy, referred to as a Hyco-entropy-based decision tree (HEDT), was created for optimum feature selection. 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 are predicated on the principle of information gain, which is efficacious for categorical variables, but its applicability to continuous variables may be unreliable. The HEDT approach overcomes this constraint by integrating entropy-based techniques to manage continuous variables. In addition, a fuzzy classifier was used for a classification analysis in order to determine the number of overlapping bacilli. Perhaps the most significant addition of this study is that it was the first to propose an automated method for bacilli counting, as opposed to previous research that accomplished this manually.

In one of the most recent publications on this subject, Vente et al. suggested a somewhat complex 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 [58]. The authors reported a 6.5% error on the test set. A second recent publication employed a multi-stage pipeline to provide a bacterial count from an image collection [63]. The pipeline was comprised of four stages: annotation using cycle-consistent generative adversarial networks (Cycle-GANs), the extraction of salient image patches, the classification of the extracted patches, and regression to obtain the final bacteria count. At every stage of the pipeline, work was performed using some kind of CNN architecture. The authors reported an error rate of less than 5% when determining the number of bacteria.

8. Discussion

We described progress in approaches to TB–AI using conventional ML methods in Section 5, DL methods in Section 6, and techniques for bacterial counting and quantification in Section 7. Furthermore, we also illustrated challenges to conducting successful research in this field. Even meticulous slide preparation for smear microscopy can generate images of variable quality with unpredictable artefacts and background staining in the sputum matrix. TB–AI works that selected only high-quality images for analysis sometimes reported significant performance results, which could not be replicated in a real-world setting. The decision-making process involved in reading stained sputum smears is inherently subjective. 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 input data. TB–AI analysis will, at least, apply the same uniform approach to the reading of 'difficult' slides.

The development of clear guidelines for the creation of image datasets to be used for TB–AI work would be beneficial. Although it may be extremely difficult for laboratories around the world to settle on completely unified approaches using identical equipment, closer agreement on the 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. The involvement of the 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 [50]. The establishment of a standardised dataset and evaluation metrics would enable researchers to evaluate the effectiveness of their methods across multiple image sets. This would reduce the necessity of adjusting model parameters and increase the applicability and comparability of the methods. For example, Zachariou et al. [63] have developed a technique that lowers the complexity of FOVs from fluorescence microscopy from coloured to greyscale images and increases the visibility of Mtb bacteria. Applying that approach to images from other datasets, including those based on brightfield images, would be useful.

Consensus agreement on commonly reported benchmarks for evaluation metrics such as the classification and segmentation of FOVs to localise Mtb bacteria would also help to reduce variability when developing and accessing new techniques. Coordination between groups active in this field, perhaps supported by WHO guidelines, may be useful here because, at present, different works using entirely different methods [63,66] to report the effectiveness of their AI methods cannot be directly compared with one another. In this work, we have successfully compiled the predominant evaluation metrics utilised in each specific category, namely classification, regression, and segmentation. As evidenced by the tabulated results presented in this paper, a distinct disparity exists among the various methods in terms of the reported evaluation metrics. Hence, if one method employs specificity as its evaluation metric, while another method incorporates both accuracy and sensitivity, it is not possible to directly compare the two methods. The presence of a substantial number of methods that are not directly comparable gives rise to a significant gap in research within this particular field, as it hampers researchers' ability to ascertain the effectiveness or ineffectiveness of these methods.

As observed, ML techniques exhibit a broad spectrum of sensitivity/recall and specificity scores. The successful integration of ML and heuristic knowledge, specifically the incorporation of anticipated cell geometric features into the algorithm, is a significant contributing factor to some methods that 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 false positives, thereby adversely impacting specificity [34,38]. 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 such as an SVM, consistently attained superior results [57,60]. In addition, akin to ML methodologies, DL techniques frequently employ amalgamated shape descriptors in the form of an additional CNN [62] or an image processing algorithm such as HOG, SURF, or CAT [54,61]. In light of the respective advantages of ML and DL, it is advisable for researchers in the domain of TB–AI to reconsider their endeavours pertaining to medical diagnosis. Sputum smear microscopy continues to hold significance, although its accuracy is contingent upon the performance of the operator in a subjective manner. Nevertheless, at present, it serves as the sole means by which microbiologists can facilitate clinical research concerning TB treatment. Consequently, a significant research gap exists in the automation of academic research associated with the monitoring of TB treatment. This is primarily due to the predominant focus of existing methods on cell detection and medical diagnosis.

Unfortunately, the current literature on the quantification of Mtb bacteria in smear microscopy images is too sparse to draw conclusions on the most appropriate methods for this aim. Certain works have utilised a pipeline approach to achieve complete automation of the quantification process, which may involve a segmentation stage. However, to improve our knowledge of the best ways to achieve bacterial quantification, additional work is necessary. At present, a limited number of methods have automated the counting process; otherwise, the counting process is carried out manually. Another manual process that has received limited attention is the creation of FOVs from sputum sample slides. As previously elucidated, the operator must perform the sequential scanning, zooming, and cropping of potential areas of the slide that may contain TB bacteria in order to make any diagnostic assessment. Hence, it is generally observed that microbiologists prioritise certain topics as being of utmost significance, while researchers focusing on TB–AI tend to allocate their efforts towards alternative areas.

Another area in which TB–AI falls behind is its deficiency in explainable artificial intelligence (XAI) techniques. Despite receiving considerable attention in multiple fields, including in healthcare and medical research [76], DL algorithms have not been widely implemented in clinical practice [77]. 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. Furthermore, 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. Finally, the model should provide justifications for its decisions by emphasising the relevant features or patterns in the input data that influenced the outcome. This is crucial for establishing the model's trustworthiness and accountability as it aligns with the overarching objective of addressing the aforementioned issue of exclusion [77]. Within the realm of tuberculosis (TB) research, XAI techniques can be employed to offer valuable understanding regarding the decision-making mechanisms of AI models utilised for diverse undertakings, including TB identification, classification, and prognosis. Several XAI techniques have the potential to be employed in TB–AI research. The integration of microscopy and XAI techniques has been explored in prior methods despite its relatively limited prevalence in the field and has proven to be highly effective in the detection of leukaemia and babesia [78,79]. The aforementioned advantages are evident; thus, it is recommended that future research efforts in the field of TB–AI incorporate these advantages as well.

9. Conclusions

Efforts to automate the analysis of sputum smear microscopy images have gradually advanced over a period over more than twenty years, 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. Image-sets used by different research groups vary because of differences in sample preparation, microscopy protocols, and imaging techniques. The absence of uniformity in datasets is important because their influence on the reported efficacy of methods has been observed to be substantial. Additionally, it would be beneficial to establish benchmarks for the evaluation of each category of TB–AI activity (namely classification, regression, and segmentation), so that work carried out by different researchers can be compared, even if those researchers also choose to employ their own additional metrics.

Notwithstanding these challenges, it may be observed that machine learning and deep learning techniques have achieved notable successes, with each approach possessing its own strengths. It has also been demonstrated in prior research that the detection of Mtb bacteria necessitates reliance on pixel intensity and shape regardless of whether the approach employs machine learning or deep learning techniques. One limitation of this work pertains to its exclusive focus on microscopy as a modality for investigating this particular disease. Several other contributions in the field of TB–AI that involve the analysis of computed tomography scans or chest radiographs from patients with pulmonary TB

have been excluded as a result of this. Some of these approached 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 certain modifications to account for their unique characteristics. In summary, in this paper we have achieved the following:

- A collection of publicly available datasets has been curated, encompassing relevant extracted data along with any supplementary annotations.
- We conclude that the provision of guidelines for both datasets and evaluation metrics is crucial in establishing standardisation. This will enable researchers to universally compare and assess their approaches.
- We have conducted a comprehensive review of existing DL/ML methods on TB–AI, specifically focusing on their application in medical diagnosis, cell detection, and cell quantification. Furthermore, we have critically examined the merits and limitations of these methods.

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 it is brightfield or fluorescence, cannot be considered the definitive standard for tuberculosis diagnosis [80–82]. However, this does not mean that TB–AI work to automate smear microscopy image analysis is no longer of value. Smear microscopy still plays an important role in assessing disease severity and monitoring therapy. The direct visualisation, quantification, and description of Mtb cells are still essential research techniques. Extending TB–AI work to count and objectively report on the phenotypic characteristics of Mtb cells during antbiotic exposure may be an important future direction for this field.

Author Contributions: Conceptualization, M.Z., O.A. and D.J.S.; methodology, M.Z.; software, M.Z.; validation, M.Z., O.A. and D.J.S.; formal analysis, M.Z., O.A. and D.J.S.; investigation, M.Z.; resources, M.Z.; data curation, M.Z.; writing—original draft preparation, M.Z.; writing—review and editing, M.Z., O.A. and D.J.S.; visualization, M.Z.; supervision, O.A. and D.J.S.; project administration, O.A. and D.J.S.; funding acquisition, D.J.S. All authors have read and agreed to the published version of the manuscript.

Funding: The authors of this paper were supported by the Welcome Trust Institutional Strategic Support fund of the University of St Andrews under grant number 204821/Z/16/Z.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Holmes, C.B.; Hausler, H.; Nunn, P. A review of sex differences in the epidemiology of tuberculosis. *Int. J. Tuberc. Lung Dis.* **1998**, *2*, 96–104. [PubMed]
2. World Health Organization. *Global Tuberculosis Report*; Technical Report; WHO: Geneva, Switzerland, 2022.
3. Kant, S.; Srivastava, M.M. Towards Automated Tuberculosis detection using Deep Learning. In Proceedings of the IEEE Symposium Series on Computational Intelligence, Xiamen, China, 6–9 December 2019; pp. 1250–1253.
4. Spence, D.P.; Hotchkiss, J.; Williams, C.S.; Davies, P.D. Tuberculosis and poverty. *Br. Med. J.* **1993**, *307*, 759–761. [CrossRef] [PubMed]
5. Gele, A.A.; Bjune, G.; Abebe, F. Pastoralism and delay in diagnosis of TB in Ethiopia. *BMC Public Health* **2009**, *9*, 5. [CrossRef]
6. Peter, J.G.; van Zyl-Smit, R.N.; Denkinger, C.M.; Pai, M. Diagnosis of TB: State of the art. *Eur. Respir. Monogr.* **2012**, *58*, 123–143.
7. Nijiati, M.; Ma, J.; Hu, C.; Tuersun, A.; Abulizi, A.; Kelimu, A.; Zhang, D.; Li, G.; Zou, X. Artificial intelligence assisting the early detection of active pulmonary tuberculosis from chest X-rays: A population-based study. *Front. Mol. Biosci.* **2022**, *9*, 874475. [CrossRef] [PubMed]
8. Chung, E.S.; Johnson, W.C.; Aldridge, B.B. Types and functions of heterogeneity in mycobacteria. *Nat. Rev. Microbiol.* **2022**, *20*, 529–541. [CrossRef]
9. Van Teeffelen, S.; Shaevitz, J.W.; Gitai, Z. Image analysis in fluorescence microscopy: Bacterial dynamics as a case study. *Bioessays* **2012**, *34*, 427–436. [CrossRef] [PubMed]
10. Ashdown, G.W.; Dimon, M.; Fan, M.; Sánchez-Román Terán, F.; Witmer, K.; Gaboriau, D.C.A.; Armstrong, Z.; Ando, D.M.; Baum, J. A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens. *Sci. Adv.* **2020**, *6*, eaba9338. [CrossRef]

11. Boeck, L.; Burbaud, S.; Skwark, M.; Pearson, W.H.; Sangen, J.; Wuest, A.W.; Marshall, E.K.P.; Weimann, A.; Everall, I.; Bryant, J.M.; et al. Mycobacterium abscessus pathogenesis identified by phenogenomic analyses. *Nat. Rev. Microbiol.* **2022**, *7*, 1431–1441. [CrossRef]
12. Toman, K. *Toman's Tuberculosis: Case Detection, Treatment and Monitoring. Questions and Answers*; World Health Organization: Geneva, Switzerland, 2004.
13. Zou, Y.; Bu, H.; Guo, L.; Liu, Y.; He, J.; Feng, X. Staining with two observational methods for the diagnosis of tuberculous meningitis. *Exp. Ther. Med.* **2016**, *12*, 3934–3940. [CrossRef]
14. Ufimtseva, E.; Eremeeva, N.; Vakhrusheva, D.; Skornyakov, S. Mycobacterium tuberculosis shape and size variations in alveolar macrophages of tuberculosis patients. *Eur. Respir. J.* **2019**, *54*, PA4605.
15. Mehta, P.K.; Raj, A.; Singh, N.; Khuller, G.K. Diagnosis of extrapulmonary tuberculosis by PCR. *FEMS Immunol. Med. Microbiol.* **2012**, *66*, 20–36. [CrossRef] [PubMed]
16. CostaFilho, C.F.F.; Levy, P.C.; Xavier, C.M.; Costa, M.G.F.; Fujimoto, L.B.M.; Salem, J. Mycobacterium tuberculosis recognition with conventional microscopy. In Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; pp. 6263–6268.
17. Steingart, K.R.; Henry, M.; Laal, S.; Hopewell, P.C.; Ramsay, A.; Menzies, D.; Cunningham, J.; Weldingh, K.; Pai, M. A systematic review of commercial serological antibody detection tests for the diagnosis of extrapulmonary tuberculosis. *Postgrad. Med. J.* **2007**, *83*, 705–712. [CrossRef] [PubMed]
18. Barr, D.A.; Schutz, C.; Balfour, A.; Shey, M.; Kamariza, M.; Bertozzi, C.R.; de Wet, T.J.; Dinkele, R.; Ward, A.; Haigh, K.A. Serial measurement of M. tuberculosis in blood from critically-ill patients with HIV-associated tuberculosis. *EBioMedicine* **2022**, *78*, 103949. [CrossRef]
19. El-Melegy, M.; Mohamed, D.; ElMelegy, T.; Abdelrahman, M. Identification of tuberculosis bacilli in ZN-stained sputum smear images: A deep learning approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019.
20. Makkapati, V.; Agrawal, R.; Acharya, R. Segmentation and classification of tuberculosis bacilli from ZN-stained sputum smear images. In Proceedings of the International Conference on Automation Science and Engineering, Bangalore, India, 22–25 August 2009; pp. 217–220.
21. Green, B.F. Public Health Image Library (PHIL). *Bull. Med. Libr. Assoc.* **2001**, *89*, 243.
22. Uddin, S. Tuberculosis Image Dataset. 2018. Available online: https://www.kaggle.com/datasets/saife245/tuberculosis-image-datasets (accessed on 22 August 2023).
23. Costa, M.G.F.; Costa Filho, C.F.F.; Kimura, A.; Levy, P.C.; Xavier, C.M.; Fujimoto, L.B. A sputum smear microscopy image database for automatic bacilli detection in conventional microscopy. In Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 26–30 August 2014; pp. 2841–2844.
24. Shah, M.I.; Mishra, S.; Yadav, V.K.; Chauhan, A.; Sarkar, M.; Sharma, S.K.; Rout, C. Ziehl–Neelsen sputum smear microscopy image database: A resource to facilitate automated bacilli detection for tuberculosis diagnosis. *J. Med. Imaging* **2017**, *4*, 027503. [CrossRef] [PubMed]
25. Trilaksana, H.; Dwimudyari, G.N.; Agoes, A.S.; Widhyatmoko, D.B. Sputum smear images database: A resource for deep learning study based to detect Bacilli for TB diagnose. In *AIP Conference Proceedings*; AIP Publishing LLC: Melville, NY, USA, 2020; Volume 2314, p. 40013.
26. Veropoulos, K.; Learmonth, G.; Campbell, C.; Knight, B.; Simpson, J. Automated identification of tubercle bacilli in sputum: A preliminary investigation. *Anal. Quant. Cytol. Histol.* **1999**, *21*, 277–282.
27. Forero-Vargas, M.; Sroubek, F.; Alvarez-Borrego, J.; Malpica, N.; Cristóbal, G.; Santos, A.; Alcalá, L.; Desco, M.; Cohen, L. Segmentation, autofocusing, and signature extraction of tuberculosis sputum images. In Proceedings of the Photonic Devices and Algorithms for Computing IV, Seattle, WA, USA, 8–9 July 2002; Volume 4788, pp. 171–182.
28. Forero, M.; Cristobal, G.; Alvarez-Borrego, J. Automatic identification techniques of tuberculosis bacteria. In *Applications of Digital Image Processing*; SPIE: Bellingham, WA, USA, 2003; Volume 5203, pp. 71–81.
29. Forero, M.G.; Sroubek, F.; Cristóbal, G. Identification of tuberculosis bacteria based on shape and color. *Real-Time Imaging* **2004**, *10*, 251–262. [CrossRef]
30. Forero, M.G.; Cristóbal, G.; Desco, M. Automatic identification of Mycobacterium tuberculosis by Gaussian mixture models. *J. Microsc.* **2006**, *223*, 120–132. [CrossRef]
31. Sadaphal, P.; Rao, J.; Comstock, G.W.; Beg, M.F. Image processing techniques for identifying Mycobacterium tuberculosis in Ziehl-Neelsen stains. *Int. J. Tuberc. Lung Dis.* **2008**, *12*, 579–582.
32. Costa, M.G.F.; Costa Filho, C.F.F.; Sena, J.F.; Salem, J.; de Lima, M.O. Automatic identification of mycobacterium tuberculosis with conventional light microscopy. In Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society, Vancouver, BC, Canada, 20–24 August 2008; pp. 382–385.
33. Sotaquira, M.; Rueda, L.; Narvaez, R. Detection and quantification of bacilli and clusters present in sputum smear samples: A novel algorithm for pulmonary tuberculosis diagnosis. In Proceedings of the International Conference on Digital Image Processing, Bangkok, Thailand, 7–9 March 2009; pp. 117–121.
34. Khutlang, R.; Krishnan, S.; Whitelaw, A.; Douglas, T.S. Automated detection of tuberculosis in Ziehl-Neelsen-stained sputum smears using two one-class classifiers. *J. Microsc.* **2010**, *237*, 96–102. [CrossRef] [PubMed]

35. Osman, M.K.; Ahmad, F.; Saad, Z.; Mashor, M.Y.; Jaafar, H. A genetic algorithm-neural network approach for Mycobacterium tuberculosis detection in Ziehl-Neelsen stained tissue slide images. In Proceedings of the International Conference on Intelligent Systems Design and Applications, Cairo, Egypt, 29 November–1 December 2010; pp. 1229–1234.
36. Osman, M.K.; Mashor, M.Y.; Saad, Z.; Jaafar, H. Colour image segmentation of tuberculosis bacilli in Ziehl-Neelsen-stained tissue images using moving k-mean clustering procedure. In Proceedings of the Asia International Conference on Mathematical/Analytical Modelling and Computer Simulation, Kota Kinabalu, Malaysia, 26–28 May 2010; pp. 215–220.
37. Osman, M.K.; Mashor, M.Y.; Jaafar, H. Detection of mycobacterium tuberculosis in Ziehl-Neelsen stained tissue images using Zernike moments and hybrid multilayered perceptron network. In Proceedings of the International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, 10–13 October 2010; pp. 4049–4055.
38. Zhai, Y.; Liu, Y.; Zhou, D.; Liu, S. Automatic identification of mycobacterium tuberculosis from ZN-stained sputum smear: Algorithm and system design. In Proceedings of the IEEE International Conference on Robotics and Biomimetics, Tianjin, China, 14–18 December 2010; pp. 41–46.
39. Nayak, R.; Shenoy, V.P.; Galigekere, R.R. A new algorithm for automatic assessment of the degree of TB-infection using images of ZN-stained sputum smear. In Proceedings of the International Conference on Systems in Medicine and Biology, Kharagpur, India, 16–18 December 2010; pp. 294–299.
40. Chang, J.; Arbeláez, P.; Switz, N.; Reber, C.; Tapley, A.; Davis, J.L.; Cattamanchi, A.; Fletcher, D.; Malik, J. 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)*; Springer: Berlin/Heidelberg, Germany, 2012.
41. Costa Filho, C.F.F.; Costa, M.G.F.; Júnior, A.K. Autofocus functions for tuberculosis diagnosis with conventional sputum smear microscopy. *Curr. Microsc. Contrib. Adv. Sci. Technol.* **2012**, *1*, 13–20.
42. Santiago-Mozos, R.; Pérez-Cruz, F.; Madden, M.G.; Artés-Rodríguez, A. An automated screening system for tuberculosis. *J. Biomed. Health Inform.* **2013**, *18*, 855–862. [CrossRef] [PubMed]
43. Ayas, S.; Ekinci, M. Random forest-based tuberculosis bacteria classification in images of ZN-stained sputum smear samples. *Signal Image Video Process.* **2014**, *8*, 49–61. [CrossRef]
44. Costa Filho, C.F.F.; Levy, P.C.; Xavier, C.d.M.; Fujimoto, L.B.M.; Costa, M.G.F. Automatic identification of tuberculosis mycobacterium. *Res. Biomed. Eng.* **2015**, *31*, 33–43. [CrossRef]
45. Govindan, L.; Padmasini, N.; Yacin, M. Automated tuberculosis screening using Zeihl Neelson image. In Proceedings of the International Conference on Engineering and Technology, Taipei, Taiwan, 22–24 April 2015; pp. 1–4.
46. Ghosh, P.; Bhattacharjee, D.; Nasipuri, M. A hybrid approach to diagnosis of tuberculosis from sputum. In Proceedings of the International Conference on Electrical, Electronics, and Optimization Techniques, Chennai, India, 3–5 March 2016; pp. 771–776.
47. Priya, E.; Srinivasan, S. Automated object and image level classification of TB images using support vector neural network classifier. *Biocybern. Biomed. Eng.* **2016**, *36*, 670–678. [CrossRef]
48. Soans, R.S.; Shenoy, V.P.; Galigekere, R.R. Automatic assessment of the degree of TB-infection using images of ZN-stained sputum smear: New results. In Proceedings of the International Conference on Systems in Medicine and Biology, Kharagpur, India, 4–7 January 2016; pp. 22–25.
49. López, Y.; Costa Filho, C.F.F.; Aguilera, L.M.R.; Costa, M.G.F. Automatic classification of light field smear microscopy patches using Convolutional Neural Networks for identifying Mycobacterium Tuberculosis. In Proceedings of the Chilean Conference on Electrical, Electronics Engineering, Information and Communication Technologies, Pucon, Chile, 18–20 October 2017; pp. 1–5.
50. Yan, S.; Liu, H.; Sun, L.; Zhou, M.; Xiao, Z.; Zhuang, Q. Detection of Mycobacterium Tuberculosis in Ziehl-Neelsen Sputum Smear Images. In Proceedings of the International Congress on Image and Signal Processing, BioMedical Engineering and Informatics, Beijing, China, 13–15 October 2018; pp. 1–6.
51. Panicker, R.O.; Kalmady, K.S.; Rajan, J.; Sabu, M.K. Automatic detection of tuberculosis bacilli from microscopic sputum smear images using deep learning methods. *Biocybern. Biomed. Eng.* **2018**, *38*, 691–699. [CrossRef]
52. Dinesh Jackson Samuel, R.; Rajesh Kanna, B. Tuberculosis (TB) detection system using deep neural networks. *Neural Comput. Appl.* **2018**, *31*, 1533–1545. [CrossRef]
53. Xiong, Y.; Ba, X.; Hou, A.; Zhang, K.; Chen, L.; Li, T. Automatic detection of mycobacterium tuberculosis using artificial intelligence. *J. Thorac. Dis.* **2018**, *10*, 1936–1940. [CrossRef]
54. Mithra, K.S.; Sam Emmanuel, W.R. FHDT: Fuzzy and Hyco-entropy-based decision tree classifier for tuberculosis diagnosis from sputum images. *Sādhanā* **2018**, *43*, 1–15. [CrossRef]
55. Díaz-Huerta, J.L.; Téllez-Anguiano, A.d.C.; Fraga-Aguilar, M.; Gutierrez-Gnecchi, J.A.; Arellano-Calderón, S. Image processing for AFB segmentation in bacilloscopies of pulmonary tuberculosis diagnosis. *PLoS ONE* **2019**, *14*, e0218861. [CrossRef] [PubMed]
56. Ahmed, T.; Wahid, F.; Hasan, J. Combining deep convolutional neural network with support vector machine to classify microscopic bacteria images. In Proceedings of the International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, Bangladesh, 7–9 February 2019; pp. 1–5.
57. Hu, M.; Liu, Y.; Zhang, Y.; Guan, T.; He, Y. Automatic detection of tuberculosis bacilli in sputum smear scans based on subgraph classification. In Proceedings of the International Conference on Medical Imaging Physics and Engineering, Shenzhen, China, 22–24 November 2019; pp. 1–7.

58. Vente, D.; Arandjelović, O.; Baron, V.O.; Dombay, E.; Gillespie, S.H. Using Machine Learning for Automatic Estimation of *M. Smegmatis* Cell Count from Fluorescence Microscopy Images. In *International Workshop on Health Intelligence*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 57–68.
59. Yousefi, H.; Mohammadi, F.; Mirian, N.; Amini, N. Tuberculosis bacilli identification: A novel feature extraction approach via statistical shape and color models. In Proceedings of the IEEE International Conference on Machine Learning and Applications, Miami, FL, USA, 14–17 December 2020; pp. 366–371.
60. Serrão, M.K.M.; Costa, M.G.F.; Fujimoto, L.B.; Ogusku, M.M.; Costa Filho, C.F.F. Automatic bacillus detection in light field microscopy images using convolutional neural networks and mosaic imaging approach. In Proceedings of the International Conference of the IEEE Engineering in Medicine & Biology Society, Montreal, QC, Canada, 20–24 July 2020; pp. 1903–1906.
61. Swetha, K.; Sankaragomathi, B.; Thangamalar, J.B. Convolutional neural network based automated detection of mycobacterium bacillus from sputum images. In Proceedings of the International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 26–28 February 2020; pp. 293–300.
62. Zachariou, M.; Arandjelović, O.; Dombay, E.; Sabiiti, W.; Mtafya, B.; Sloan, D. Extracting and Classifying Salient Fields of View From Microscopy Slides of Tuberculosis Bacteria. In Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence, Xiamen, China, 23–25 September 2022; pp. 1–12.
63. Zachariou, M.; Arandjelović, O.; Sabiiti, W.; Mtafya, B.; Sloan, D. Tuberculosis bacteria detection and counting in fluorescence microscopy images using a multi-stage deep learning pipeline. *Information* **2022**, *13*, 96. [CrossRef]
64. Hicks, S.A.; Strümke, I.; Thambawita, V.; Hammou, M.; Riegler, M.A.; Halvorsen, P.; Parasa, S. On evaluation metrics for medical applications of artificial intelligence. *Sci. Rep.* **2022**, *12*, 5979. [CrossRef] [PubMed]
65. Chalana, V.; Kim, Y. A methodology for evaluation of boundary detection algorithms on medical images. *IEEE Trans. Med. Imaging* **1997**, *16*, 642–652. [CrossRef]
66. Khutlang, R.; Krishnan, S.; Dendere, R.; Whitelaw, A.; Veropoulos, K.; Learmonth, G.; Douglas, T.S. Classification of Mycobacterium tuberculosis in images of ZN-stained sputum smears. *Trans. Inf. Technol. Biomed.* **2010**, *14*, 949–957. [CrossRef]
67. Ayma, V.; De Lamare, R.; Castañeda, B. An adaptive filtering approach for segmentation of tuberculosis bacteria in Ziehl-Neelsen sputum stained images. In Proceedings of the Latin America Congress on Computational Intelligence, Curitiba, Brazil, 13–16 October 2015; pp. 1–5.
68. Frie, T.T.; Cristianini, N.; Campbell, C. The kernel-adatron algorithm: A fast and simple learning procedure for support vector machines. In Proceedings of the Machine Learning: Proceedings of the Fifteenth International Conference, Madison, WI, USA, 24–27 July 1998; pp. 188–196.
69. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
70. Mithra, K.S.; Sam Emmanuel, W.R. Automated identification of mycobacterium bacillus from sputum images for tuberculosis diagnosis. *Signal Image Video Process.* **2019**, *13*, 1585–1592. [CrossRef]
71. He, D.; Xia, Y.; Qin, T.; Wang, L.; Yu, N.; Liu, T.Y.; Ma, W.Y. Dual learning for machine translation. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 820–828.
72. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
73. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
74. Imperial, M.Z.; Nahid, P.; Phillips, P.P.J.; Davies, G.R.; Fielding, K.; Hanna, D.; Hermann, D.; Wallis, R.S.; Johnson, J.L.; Lienhardt, C. A patient-level pooled analysis of treatment-shortening regimens for drug-susceptible pulmonary tuberculosis. *Nat. Med.* **2018**, *24*, 1708–1715. [CrossRef]
75. Payasi, Y.; Patidar, S. Diagnosis and counting of tuberculosis bacilli using digital image processing. In Proceedings of the International Conference on Information, Communication, Instrumentation and Control, Indore, India, 17–19 August 2017; pp. 1–5.
76. Cooper, J.; Arandjelović, O.; Harrison, D.J. Believe the HiPe: Hierarchical perturbation for fast, robust, and model-agnostic saliency mapping. *Pattern Recognit.* **2022**, *129*, 108743. [CrossRef]
77. de Vries, B.M.; Zwezerijnen, G.J.C.; Burchell, G.L.; van Velden, F.H.P.; Boellaard, R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: A literature review. *Front. Med.* **2023**, *10*, 1180773. [CrossRef]
78. Abir, W.H.; Uddin, M.F.; Khanam, F.R.; Tazin, T.; Khan, M.M.; Masud, M.; Aljahdali, S. Explainable AI in diagnosing and anticipating leukemia using transfer learning method. *Comput. Intell. Neurosci.* **2022**, *2022*, 5140148. [CrossRef] [PubMed]
79. Durant, T.J.S.; Dudgeon, S.N.; McPadden, J.; Simpson, A.; Price, N.; Schulz, W.L.; Torres, R.; Olson, E.M. Applications of digital microscopy and densely connected convolutional neural networks for automated quantification of babesia-infected erythrocytes. *Clin. Chem.* **2022**, *68*, 218–229. [CrossRef] [PubMed]
80. Boldi, M.O.; Denis-Lessard, J.; Neziri, R.; Brouillet, R.; Von-Garnier, C.; Chavez, V.; Mazza-Stalder, J.; Jaton, K.; Greub, G.; Opota, O. Performance of microbiological tests for tuberculosis diagnostic according to the type of respiratory specimen: A 10-year retrospective study. *Front. Cell. Infect. Microbiol.* **2023**, *13*, 1131241. [CrossRef] [PubMed]

81. van Dijk, S.G.; Scheunemann, M.M. Deep learning for semantic segmentation on minimal hardware. In *Robot World Cup*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 349–361.
82. Palomino, J.C. Nonconventional and new methods in the diagnosis of tuberculosis: Feasibility and applicability in the field. *Eur. Respir. J.* **2005**, *26*, 339–350. [CrossRef] [PubMed]