JOURNAL OF
EMERGING INVESTIGATORS

Article

Effects of data amount and variation in deep learning-based tuberculosis diagnosis in chest X-ray scans

Anusuiya Bhorkar[1], **Ricardo A. Gonzales**[2]

[1] Dominican Academy, New York City, New York

[2] Radcliffe Department of Medicine, University of Oxford, Oxford, United Kingdom

SUMMARY

Pulmonary tuberculosis ranks among the world's deadliest diseases, causing devastating global health harm. Despite diagnosing millions annually, COVID-19's aftermath and inadequate screening within developing countries interfered with diagnosis of tuberculosis, hindering proper treatment and increasing tuberculosis mortality. Our study aims to enhance diagnostic opportunities by developing a deep-learning model to categorize pulmonary X-ray scans into "normal" and "tuberculosis" classes. Exercising feature extraction, we incorporated VGG-16 (Visual Geometry Group network with 16 layers) architecture to enhance model classification accuracy via transfer learning. We hypothesized that models trained on a greater amount and variety of data would perform better than those trained on less invariable data. Testing our hypothesis, we developed four models with replicate architectures trained on pulmonary X-ray scans from the Montgomery County Chest X-ray Dataset (Dataset A), the Shenzhen Chest X-ray Dataset (Dataset B), the Tuberculosis Chest X-ray Database (Dataset C), and the combined data (Dataset A, B, and C). Data amount and variability increased from Dataset A to C and was the largest in the combined datasets. Testing these models on each dataset used in training, we found the mean accuracy values were 50.3% for Dataset A, 57.9% for Dataset B, 79.3% for Dataset C, and 98.9% for the combined dataset. This indicates that models trained on more data perform more accurately across datasets due to greater data variation and amount. Additionally, our study highlights the efficacy of deep learning models in tuberculosis diagnosis, emphasizing the importance of data variability and a wider pulmonary X-ray database for potential implementation in global health.

INTRODUCTION

The tuberculosis illness, caused by the bacteria *Mycobacterium tuberculosis*, afflicts the lives of millions today. The disease mainly affects the lungs, making pulmonary tuberculosis one of the most common presentations of tuberculosis (1). According to the 2023 report by the World Health Organization (WHO), 7.5 million new tuberculosis cases were diagnosed in 2022, marking the highest annual count since the WHO started tracking the disease globally in 1995 (2). Yet, since the COVID-19 pandemic in 2020, the number of people receiving diagnoses has decreased due to limited access to healthcare services since technical and financial resources were mainly directed to the COVID-19 outbreak. This continues to relate today because the number of people who received tuberculosis diagnoses declined by 18% in 2020 as compared to 2019, while the global number of tuberculosis deaths between 2019 and 2023 grew from 1.2 million to 1.3 million deaths per year, causing the WHO plan to eradicate tuberculosis to be slowed by a decade (3). People cannot be properly treated without receiving a proper diagnosis, causing their illnesses to develop further. This issue is more urgently pressing with the increase in drug-resistant and antibiotic-resistant tuberculosis strains that require quicker treatment and are more invulnerable to common treatment methods (4). Additionally, the 2022 WHO report shows that the majority of tuberculosis-related deaths occur in developing countries that have a scarcity of medical screening initiatives (5). This establishes a direct correlation between a lack of tuberculosis diagnosis and an increase in tuberculosis-caused deaths (4). In 2022, the death rate from tuberculosis without treatment was about 50%, while 85% of those receiving treatment were predicted to be cured (6).

Despite the setbacks from COVID-19 and the limitations of developing countries, diagnoses are still being given to patients in need (7). Nonetheless, the methods used to diagnose patients leave room for improvement. For example, direct sputum smear microscopy is the most commonly used method for diagnosing pulmonary tuberculosis, but it can be costly and inconvenient for patients as it requires multiple visits to healthcare facilities and the collection of several sputum samples over a period of days (8). There are multiple methods of smear microscopy testing, with the most common test being Ziehl-Neelsen-stained smears. Despite its high specificity, its sensitivity shows variability as it ranges from 20%-80% (9). This means that although this method has a low rate of false positives, its ability to detect true positives varies. Nucleic acid amplification tests, or NAA tests, are also used as a method of diagnosis; however, their sensitivity is also too poor to rule out diseases, and the Food and Drug Administration recommends a secondary sputum test afterward (10). This increases patient wait time for a diagnosis. Pulmonary X-ray analysis is also highly considered when diagnosing tuberculosis and is continuously used in treatment as there are similarities often found among tuberculosis-positive chest X-rays. However, pulmonary chest X-ray analysis demands attention and reading time from healthcare professionals, demonstrating variability and bias within their diagnoses (11). Therefore, resulting from the cost-wise and time-wise inconvenience and lack of accuracy in current tuberculosis diagnosis, further research is required to develop a more precise and efficient method of optimizing patient diagnosis for tuberculosis, allowing for more patients to have access to correct diagnoses to receive proper treatment.

Recently, the emerging field of deep learning has gained attention in medicine, particularly through its application to medical image analysis. The ability of Artificial Intelligence (AI) to classify medical images shows the potential to improve diagnosis accuracy, cost, and speed (12, 13). Specifically, the VGG-16 (Visual Geometry Group network with 16 layers) model has shown great effectiveness, reaching a top-5 test accuracy of 92.7% on ImageNet, a dataset consisting of 14 million images across 1,000 categories (14, 15). Accordingly, the VGG-16 model can be adapted for use in analyzing pulmonary X-rays and labeling the data that has been classified as showing signs of tuberculosis or as "normal" in a manner that is both efficient, inexpensive, and accurate. Although there are other available model architectures to choose from, we have chosen the VGG-16 architecture because, unlike other model architectures, it does not necessitate reliance on a large number of hyperparameters (model settings) or additional computational cost in order to maintain its effectiveness. Additionally, due to its repetitiveness within its pattern of convolutional blocks (groupings of layers that extract features from the data), it has much stronger feature-learning abilities than other architectures, such as AlexNet (14). Yet, when constructing such a model, the architecture and training data both have a sizeable impact on the model's results. The model's accuracy may be different depending on the different amounts of data and data variability. A VGG-16 model with a large amount of variable data could be developed to act as an additional reviewer to greatly assist healthcare workers in interpreting pulmonary radiography data, as it allows for more patients to be diagnosed in a shortened amount of time. Transfer learning, a technique where a pre-trained model is fine-tuned for a specific task by leveraging knowledge from a broader dataset, could further enhance this model by improving its ability to identify tuberculosis-related features efficiently and accurately.

Further adapting this architecture via transfer learning, in this work, we aimed to utilize deep learning to form a more accurate and efficient method of diagnosing patients with pulmonary tuberculosis while investigating the impact of different data amounts and variability. We achieved this by developing four deep-learning models using the same VGG-16 architecture via transfer learning. With this architecture, the models were trained to classify pulmonary X-ray scans into two classes: "tuberculosis" and "normal". Yet, while the models had the same architecture, they were trained on different databases of pulmonary X-ray scans from the Montgomery County (Dataset A), the Shenzhen Set (Dataset B), the Tuberculosis Chest X-ray Database (Dataset C), and all datasets combined (Datasets A, B, and C). The models were then tested on each dataset in addition to a combination of the three datasets. As the amount of data and variety of data of each model increased from Dataset A to Dataset C, with the combined dataset consisting of the most data and variety, the testing results allowed us to visualize how the different data amounts and variability within the data impacts the accuracy with which the deep learning model is able to classify the pulmonary X-ray scans. We hypothesized that the models trained on larger and more variable data would perform with a higher accuracy rate than models trained on smaller and less variable data. This hypothesis was supported by our findings, as the model trained on Dataset C–being the largest and most varied individual dataset–achieved the highest accuracy among models trained on single datasets. Additionally, the model trained on the combined dataset, which had the most variability and the largest amount of data overall, achieved the best performance in the study.

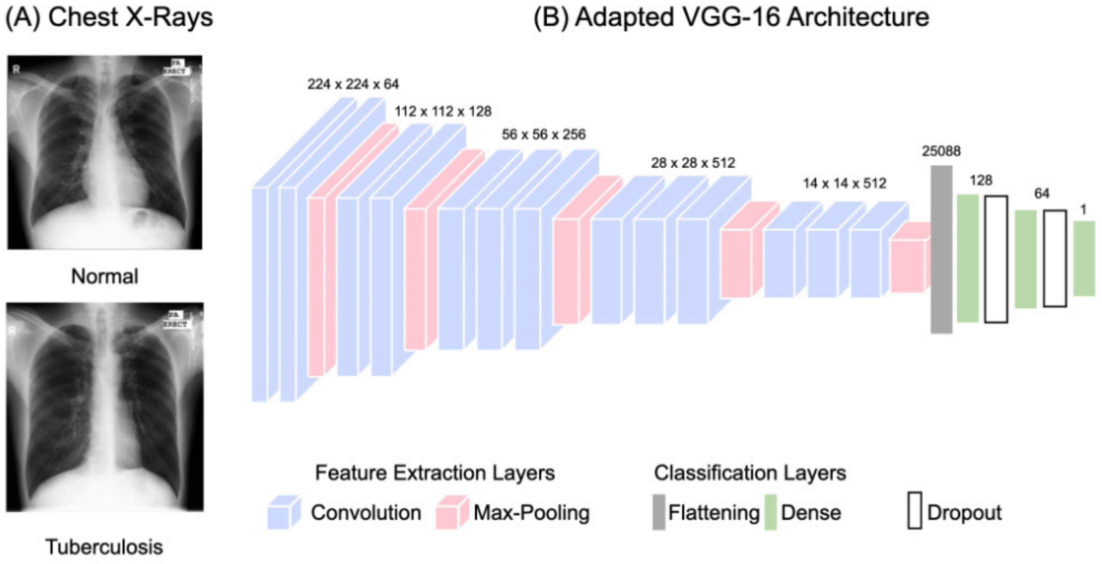

Figure 1: Imaging data and deep learning architecture. (A) Two images showcase pulmonary X-ray scans. One represents a normal scan (top) and the other represents a scan showing tuberculosis (bottom). (B) Model showcasing the layers involved in the structure of the VGG-16 architecture (14). The model's architecture follows a funnel method, consisting of 13 convolutional layers grouped into five blocks, where the numbers within each block represent the channels within each layer.

https://doi.org/10.59720/24-278

RESULTS

In our study, we developed four models using four separate datasets consisting of pulmonary X-ray scans testing both positively and negatively for tuberculosis: Datasets A, B, C, and combined (A + B + C) (**Figure 1A**). The models each followed identical VGG-16 architecture (**Figure 1B**). Each model was trained on one of the above datasets and tested on all four to compare the accuracies of properly identifying tuberculosis cases (**Table 1**). For clarity, each accuracy level was categorized by performance range and color: "low" (under 50%) in red, "moderate" (50% to 75%) in yellow, and "high" (75% to 100%) in green. Additionally, we created confusion matrices to visualize how the data was organized when each model was further tested on the same dataset it was trained on (**Figure 2**).

The confusion matrices revealed that the models tested on the dataset they were trained on scored in the "high" range. In this category, the model trained on Dataset A reached 89.3% accuracy, the model trained on Dataset B reached 86.8% accuracy, the model trained on Dataset C reached 95.9% accuracy, and the model trained on Dataset A + B + C reached 97.9% accuracy. However, when the models were tested on data from the other datasets, there was more variability. The model trained on Dataset A reached testing accuracies of 52.6%, 27.2%, and 32.2%, tested on Datasets B, C, and the combined dataset, respectively. The model trained on Dataset B reached testing accuracies of 54.5%, 39.1%, and 51.0% when tested on Datasets A, C, and the combined dataset, respectively. The model trained on Dataset C reached testing accuracies of 56.6%, 69.1%, and 95.4%, tested on Datasets A, B, and the combined dataset, respectively. Finally, the model trained on the combined dataset reached a value of 100.0%, 98.1%, and 99.6%, tested on Datasets A, B, and C, respectively. Therefore, the mean values of the models' accuracies when tested on each dataset were 50.3% for Dataset A, 57.9% for Dataset B, 79.3% for Dataset C, and 98.9% for Dataset A + B + C (**Table 1**). Each value was rounded to the nearest tenth. All-embracing, our initial hypothesis was supported by our data as the models with greater training data variability and amount outperformed the models with lesser amounts of data and variability in training, and it was found that the models performed better on data from seen datasets than they did on unseen datasets.

DISCUSSION

The overarching aim of our research was to develop deep-learning models using feature extraction that could accurately diagnose patients with pulmonary tuberculosis by classifying pulmonary X-ray scans, allowing for the exploration of the effect of increased training data variability and amount of data on a model's testing performance. Using an adapted VGG-16 architecture with transfer learning to create the four models, we discovered that the models trained on more data with greater data variability led to higher accuracies, with the model trained on all the datasets performing the best when tested on each dataset. We also found that models trained on certain datasets performed better on data with features similar to those of the training data. These findings show the potential of AI and deep learning as a second reader to facilitate tuberculosis diagnosis, in addition to displaying the importance of increasing the amount and variability of training data when creating such a model. Therefore, these findings emphasize the need for an expanded collection of pulmonary X-ray images portraying both tuberculosis and normal readings. Based on the multitude of tuberculosis cases worldwide, especially in developing countries, a deep learning model trained on widespread tuberculosis data could help serve as a second reader for healthcare practitioners in such areas to allow for the more cost and time-efficient reading of pulmonary X-ray images.

Test Train	A (n = 27)	B (n = 132)	C (n = 840)	A + B + C (n = 1000)
A (n = 111)	89.3%	52.6%	27.2%	32.2%
B (n = 530)	54.5%	86.8%	39.1%	51.0%
C (n = 3360)	56.6%	69.1%	95.9%	95.4%
A + B + C (n = 4001)	100.0%	98.1%	99.6%	97.9%

Table 1: Testing accuracies for each model across all datasets.

When analyzing the accuracy levels of the models, it is important to note what type of content the datasets consist of. Although each dataset consists of tuberculosis and normal images, their sizes and variability differ, as scans across more patients allow for different types of pulmonary systems to be addressed by the model. Additionally, the different datasets have differing normal-to-tuberculosis data ratios. Dataset A consists of 130 images in total, with a roughly 1.38:1.00 ratio of normal scans to tuberculosis scans, Dataset B consists of 662 images in total, with a roughly 0.97:1.00 ratio of normal scans to tuberculosis scans, and Dataset C consists of 4,200 images in total, with a roughly 5.00:1.00 ratio of normal scans to tuberculosis scans. The combined Dataset A + B + C consists of 4992 images in total, with a roughly 3.57:1.00 ratio of normal scans to tuberculosis scans. Dataset A is, therefore, the dataset with the least amount of data, with Dataset C having the most amount of data out of the models trained on individual datasets. The set with the most data overall is Dataset A + B + C. While Dataset C has the largest disparity between the number of normal scans and tuberculosis scans, increasing its chance of bias, it's important to note that due to a much larger number of patient scans in Dataset C, more diverse pulmonary systems are being taken into account within the dataset, thus increasing variability. Furthermore, Dataset A only has 38 more normal scans than tuberculosis scans. Due to the small size of the dataset, this amount is not enough to be valued at a high variability.

Similarly, other studies used deep learning to examine chest X-rays; for instance, a convolutional neural network called DeTraC was proposed to aid the COVID-19 diagnosis process (16). Another study aimed to explore how different tuberculosis classification model architecture affects the models' respective accuracies using the Shenzhen Set-Chest X-ray Database and Montgomery Set Chest X-ray Database (17). Within the VGG-16 model testing, the proposed model reached an accuracy of 85.74% for the Shenzhen Set Chest X-ray Database and 77.14% for the Montgomery Chest

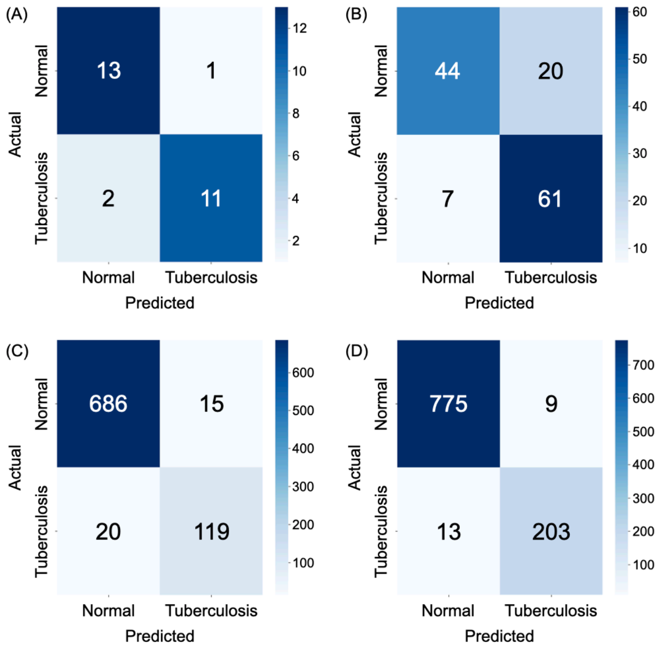

Figure 2: Confusion matrices showcasing model performance when tested on held-out data from the same dataset used for training. Each panel represents a model trained and tested on different subsets of a specific dataset: (A) Dataset A (27 test images), (B) Dataset B (132 test images), (C) Dataset C (840 test images), and (D) Combined Dataset A + B + C (1000 test images). In each matrix, the top left indicates true normal predictions, bottom right true tuberculosis, top right false tuberculosis (normal predicted as tuberculosis), and bottom left false normal (tuberculosis predicted as normal).

X-ray Database (17). Contrasting this research, in our study, the models are evaluated both internally and externally to provide a greater view of how our models perform on unseen data. Additionally, rather than exploring the use of differing architectures, our study focused on the differences in accuracy when different data was used. This allows for the results of our study to display the relationship between greater data amount and variety and increased model performance by comparing the accuracies between each model. In addition, our study explores the positive trend in model accuracy when further tested on the same data. The confusion matrices revealed that when each model was further tested using the same data it was trained on, the models all had testing accuracies in the "high" category, which provides evidence that models are better at classifying data when the data has features similar to the ones extracted from the training data. Therefore, if a model has more data with higher variability, it could extract more features from the dataset and therefore recognize more features in testing data to classify said data more accurately. The model trained on Dataset A has the lowest mean accuracy of 50.3%, with its accuracies across datasets falling into the "low" category except for the instance in which Dataset A was further tested on itself and reached an accuracy of 89.3%. This suggests that models trained on datasets with low amounts of data and low variability of data perform poorly on unseen data and well on seen data because they lack access to generalized data within training and thus become overfit to a particular dataset. Therefore, they would not be as effective in diagnosing tuberculosis patients in a real-world setting. Next, in the model trained on Dataset B, the training data has more variability and is slightly larger than Dataset A but relatively smaller in comparison to Dataset C. It also has a nearly equivalent normal-to-tuberculosis scan ratio. While this eliminates the potential bias of the model (where the model is more inclined to label an image as a tuberculosis reading rather than a normal one or vice versa), the smaller size means reduced variability despite the variability being greater than that of Dataset A. When tested, the mean accuracy of Dataset B was recorded as 57.9%. Yet, this difference in accuracies between Dataset A and Dataset B reveals that an increase in data size and variation improves the overall performance of a model, in addition to a more equivalent ratio of information within a dataset to eliminate bias and perform better on unseen data. Examining the testing accuracy of Dataset C, we observed that the mean accuracy value was 79.3%, a higher overall performance compared to Dataset A and Dataset B. This increase in accuracy also supports the claim that greater data amounts and variation positively impact the performance of a model, as Dataset C is 5.25 times larger than both Datasets A and B combined. However, the disparity between the amount of normal pulmonary scans and pulmonary scans exhibiting signs of tuberculosis may have possibly resulted in a bias, which could have brought down the testing accuracies on datasets aside from Dataset C. Yet, the sheer size of Dataset C outweighed the potential bias, as each testing accuracy for Dataset C fell in either the "high" or "medium" categories. In future research, a model of this size could be tested using an equivalent amount of "normal" and "tuberculosis" pulmonary X-ray scans, which we do not currently have access to. Even with these current results, the trend in data amount and variability increase, and increased model accuracy is observed.

In our analysis of the model trained on Dataset A + B + C, we see the greatest variety as this model consists of images across all three separate datasets and is therefore the largest out of each dataset with a data count of 5,000 images. The overall accuracy for this model is 98.9%, the highest overall accuracy out of each model. Hence, it can be determined that a high variability and amount of data strongly impact the accuracy of datasets because the model trained on the largest dataset scored the highest. It is still important to note that the model's increased accuracy was in part due to the model's training data consisting of images from each dataset, allowing us to observe that (because of feature extraction) models with training data familiar to the data they are tested on tend to perform better on such data. This supports the idea that data amount and variability affect model accuracy. A model trained on a larger and more diverse dataset encounters a greater number of unique features, which enhances its ability to generalize. As a result, the model can better recognize the characteristics of unseen data by comparing it to the extensive and varied data it has already recognized.

For future research, taking the combined dataset and testing it on yet another dataset would aid in solidifying this claim. However, we are currently restricted due to the current number of available pulmonary X-ray tuberculosis datasets. It is hence important to first collect and organize diverse pulmonary X-ray scans for the creation of a much larger pulmonary X-ray database. Furthermore, due to variations in pulmonary X-rays across ethnicities, genders, and ages, we aim to standardize data distribution across these categories to enhance the project's scope. However, we currently lack sufficient diverse data to minimize bias. Furthermore, while dropout layers were implemented in the model architecture to mitigate the over-specialization of individual neurons and decrease overfitting, our models can become additionally less prone to memorizing noise with the addition of data augmentation layers, which would help to create new training examples by applying various transformations to existing data.

In summary, our model serves to showcase the effects of data variability and data amount, supporting our hypothesis that models trained on greater data amounts with increased variety are more accurate in their performances. The results from the testing data of our four VGG-16 models indicate that models trained on less data with lesser variability are more specialized to individual data and therefore more overfit to specific datasets within training data; models trained on more training data with more variability are more generalized and have practiced extracting more features that make them more capable in classifying external data; and models trained on data from certain datasets will showcase higher performance when further tested on the same data. Therefore, the implication of this study is the demonstrated importance of the collection of greater amounts of data and more data variability within pulmonary X-ray scans of the normal and tuberculosis categories for the creation of a more accurate model that can be used to increase tuberculosis diagnosis. Such a model would better act as an additional reviewer to shorten diagnosis time by assisting healthcare workers in pulmonary radiography interpretations, increasing efficiency in proper tuberculosis diagnoses as the first step towards treatment.

MATERIALS AND METHODS

Image Datasets

The image datasets used to train and test the different versions of our model included Dataset A, Dataset B, and Dataset C (18, 19, 20). The Montgomery County - Chest X-ray Database, created by the Department of Health and Human Services in Montgomery County, Maryland, USA, contains 58 X-ray cases showing tuberculosis and 80 normal cases (18). The dataset includes 74 scans from female patients and 63 from male patients, with ages ranging from 4 to 89 years. However, the data has been de-identified to maintain patient privacy. The de-identified Shenzhen Set-Chest X-ray Database contains X-rays sourced from Shenzhen No.3 Hospital in Shenzhen, China, where 336 X-rays portray cases with manifestations of tuberculosis and 326 X-rays portray normal cases, with 205 female scans to 457 male scans and a patient age range of 1 to 89 (18,19). The TB Chest X-ray Database was developed by researchers from Qatar University in Doha, the University of Dhaka in Bangladesh, and collaborators from Malaysia, working with medical doctors from Hamad Medical Corporation and Bangladesh (20). This database includes 4,200 pulmonary X-ray images, with 3,500 labeled as "normal" and 700 as showing "tuberculosis." The gender and average age of the patients have been anonymized.

Data Preprocessing

Before training, testing, and creating the combined dataset, Dataset A and Dataset C were pre-processed to enable the model to interpret the raw data. This was not required for Dataset B because it was already in the desired format as originally published (19). The images from Dataset C were separated for storage into two separate folders, requiring two rounds of preprocessing, while Dataset A's images came from one folder and were therefore preprocessed together. The images were converted into NumPy arrays, saved as integer values, and cropped to remove padding before being saved as PIL images in a new folder with a "jpg" extension. Dataset A, which had not been split into classes, was then divided using an if-statement to evaluate whether an image represented a "normal" or "tuberculosis" X-ray scan. This was based on file names marked with "0" for normal and "1" for tuberculosis, and placed into the appropriate folder.

Model Architecture

Our model follows a convolutional neural network with adapted VGG-16 architecture to categorize the images into two classes (**Figure 1B**). A convolutional neural network (CNN) is a type of deep learning model specifically designed to analyze visual data by identifying patterns such as edges or textures. The VGG-16 model, pre-trained on the ImageNet Database, was incorporated into our model via transfer learning, a technique that repurposes a model trained on one task for a different but related task, saving time and resources while improving performance. The last three fully connected layers of the original VGG-16 were removed, and the remaining layers were used to extract features from the X-ray images. The batch input shape (the dimensions of input images) was adjusted to match our specified image size. To adapt the base model, flattening layers were added to transform multidimensional data into a single vector for processing; fully connected (dense) layers with a rectified linear activation function, a mathematical function that helps the model learn relationships between features; dropout layers, which randomly deactivate some neurons during training to prevent overfitting; and a final dense layer with sigmoid activation, which outputs a probability between 0 and 1. This probability indicates the model's confidence in classifying an input image as either tuberculosis-positive or normal. Images with probabilities of 0.5 or higher are classified as tuberculosis, while those below 0.5 are classified as normal, providing a clear diagnostic output.

Model Optimization

To increase model accuracy, we incorporated transfer learning and tuned several hyperparameters. Transfer learning aided in the development of our model as it ensured that our model would already be able to extract useful features from a large dataset, giving the model experience in capturing important patterns and structures in images that could be applied to feature extraction of pulmonary X-ray scans for our purposes. In our project, the base model that we further adapted was pre-trained using the ImageNet Database, a large visual database consisting of over 3 million varied and diverse images (15). Within hyperparameter tuning, we set our batch size to eight images per training iteration, providing a regularizing effect and a lower computational burden for our models. As for the image size, we specified the images as having a 224 × 224-pixel dimension with only one grayscale channel. Setting the seed allowed for the random processes

https://doi.org/10.59720/24-278

in the training to be the same across different runs, and the split ratio of 0.20 ensured 80% of the images would be used for training. When the models later underwent external testing and further testing on the training dataset, 20% of each dataset was used in testing (**Table 1, Figure 2**). As for the number of iterations the training datasets would undergo, 50 epochs were chosen, resulting from the computational limitations of how long the model could run, in addition to preventing the model from being underfitted and unable to perform well on the internal validation data. We used the Adam optimizer to minimize the loss function of the models. Additionally, the models employed the binary cross entropy loss function to analyze the disparity between the predicted binary outcomes and true binary labels, guiding the algorithm toward more accurate classifications. These hyperparameters influenced the behavior and performance of the models during training.

Model Implementation

The code utilized for the pre-processing, model training, and model testing was created in Google Colaboratory using the Python coding language (21). Because of its integration with Google Drive, the retrieval and organization of data from various datasets and classes were facilitated, making Google Colab an optimal choice. Additionally, we utilized Tensorflow Framework throughout the development of the models. Since the length of the time needed to train a model is dependent on the testing data size, hyperparameters, architectures, and other characteristics regarding the model, our four models performed with different speeds (22). When trained for 50 epochs, the approximate time in minutes each model took to train was 7.2 minutes for Dataset A, 5.2 minutes for Dataset B, 24.3 minutes for Dataset C, and 28.1 minutes for the combined dataset. The scripts for preprocessing the data and training and evaluating the model can be accessible at the following GitHub repository: https://github.com/AnusuiyaBhorkar/TB-Deep-Learning.

Model Evaluation

The results of our study were evaluated for their performance based on the accuracy of each model. This is dependent on the percentage of which the last dense layer of the model ultimately classifies the inputted testing data correctly. Accuracies under 50% were labeled as "low" and categorized by red, accuracies between 50% to 75% were labeled as "moderate" and categorized by yellow, and accuracies from 75% to 100% were labeled as "high" and categorized by green.

ACKNOWLEDGMENTS

We wish to acknowledge and thank Veritas AI for the support given throughout this study.

Received: September 26, 2024
Accepted: December 13, 2024
Published: April 28, 2025

REFERENCES

1. Bai, Wentao, and Edward Kwabena Ameyaw. "Global, Regional and National Trends in Tuberculosis Incidence and Main Risk Factors: A Study Using Data from 2000 to 2021." *BMC Public Health*, vol. 24, no. 1, Jan. 2024, https://doi.org/10.1186/s12889-023-17495-6.
2. "Global Tuberculosis Report 2023." *World Health Organization*. www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023. Accessed 24 May 2024.
3. Sanduzzi Zamparelli, Stefano, *et al.* "Clinical Impact of COVID-19 on Tuberculosis." *Infezioni in Medicina*, vol. 30, no. 4, Dec. 2022, https://doi.org/10.53854/liim-3004-3.
4. Jeremiah, Chakaya, *et al.* "The WHO Global Tuberculosis 2021 Report – Not so Good News and Turning the Tide back to End TB." *International Journal of Infectious Diseases*, vol. 124, no. 1, Mar. 2022, https://doi.org/10.1016/j.ijid.2022.03.011.
5. Bagcchi, Sanjeet. "WHO's Global Tuberculosis Report 2022." The Lancet Microbe, vol. 4, no. 1, Dec. 2022, https://doi.org/10.1016/S2666-5247(22)00359-7.
6. Hopewell, Philip C, *et al.* "International Standards for Tuberculosis Care." *The Lancet Infectious Diseases*, vol. 6, no. 11, Nov. 2006, pp. 710–725, https://doi.org/10.1016/s1473-3099(06)70628-4.
7. Sarınoğlu, Rabia Can, *et al.* "Tuberculosis and COVID-19: An Overlapping Situation during Pandemic." *The Journal of Infection in Developing Countries*, vol. 14, no. 07, July 2020, pp. 721–725, https://doi.org/10.3855/jidc.13152.
8. Ryu, Yon Ju. "Diagnosis of Pulmonary Tuberculosis: Recent Advances and Diagnostic Algorithms." *Tuberculosis and Respiratory Diseases*, vol. 78, no. 2, 2015, p. 64, https://doi.org/10.4046/trd.2015.78.2.64.
9. Dzodanu, Eben Godsway, *et al.* "Diagnostic Yield of Fluorescence and Ziehl-Neelsen Staining Techniques in the Diagnosis of Pulmonary Tuberculosis: A Comparative Study in a District Health Facility." *Tuberculosis Research and Treatment*, vol. 2019, Apr. 2019, https://doi.org/10.1155/2019/4091937.
10. Greco, Stefania, *et al.* "Current Evidence on Diagnostic Accuracy of Commercially Based Nucleic Acid Amplification Tests for the Diagnosis of Pulmonary Tuberculosis." *Thorax*, vol. 61, no. 9, 1 Sept. 2006, pp. 783–790, https://doi.org/10.1136/thx.2005.054908.
11. Harris, Miriam, *et al.* "A Systematic Review of the Diagnostic Accuracy of Artificial Intelligence-Based Computer Programs to Analyze Chest X-Rays for Pulmonary Tuberculosis." *PLOS ONE*, vol. 14, no. 9, Sept. 2019, p. e0221339, https://doi.org/10.1371/journal.pone.0221339.
12. Bohr, Adam, and Kaveh Memarzadeh. "The Rise of Artificial Intelligence in Healthcare Applications." *Artificial Intelligence in Healthcare*, vol. 1, no. 1, 2020, pp. 25–60. *NCBI*, https://doi.org/10.1016/B978-0-12-818438-7.00002-2.
13. Guan, Qing, *et al.* "Deep Convolutional Neural Network VGG-16 Model for Differential Diagnosing of Papillary Thyroid Carcinomas in Cytological Images: A Pilot Study." *Journal of Cancer*, vol. 10, no. 20, 2019, pp. 4876–4882, https://doi.org/10.7150/jca.28769.
14. Simonyan, Karen, and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." *ICLR 2015 - 3rd International Conference on Learning Representations*, 2015, https://doi.org/10.48550/arXiv.1409.1556.
15. Deng, Jia, *et al.* "ImageNet: A Large-Scale Hierarchical Image Database." *2009 IEEE Conference on Computer*

Vision and Pattern Recognition, June 2009, https://doi.org/10.1109/cvpr.2009.5206848.

16. Abbas, Asmaa, *et al.* “Classification of COVID-19 in Chest X-Ray Images Using DeTraC Deep Convolutional Neural Network.” *Applied Intelligence*, Sept. 2020, https://doi.org/10.1007/s10489-020-01829-7.
17. Meraj, Syeda Shaizadi, *et al.* “Detection of Pulmonary Tuberculosis Manifestation in Chest X-Rays Using Different Convolutional Neural Network (CNN) Models.” *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, Oct. 2019, pp. 2270–2275, https://doi.org/10.35940/ijeat.a2632.109119.
18. Candemir, Sema, *et al.* “Lung Segmentation in Chest Radiographs Using Anatomical Atlases with Nonrigid Registration.” *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, Feb. 2014, pp. 577–590, https://doi.org/10.1109/tmi.2013.2290491.
19. Jaeger, Stefan, *et al.* “Automatic Tuberculosis Screening Using Chest Radiographs.” *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, Feb. 2014, pp. 233–245, https://doi.org/10.1109/tmi.2013.2284099.
20. Rahman, Tawsifur, *et al.* “Reliable Tuberculosis Detection Using Chest X-Ray with Deep Learning, Segmentation and Visualization.” *IEEE Access*, vol. 8, 2020, pp. 191586–191601, https://doi.org/10.1109/access.2020.3031384.
21. Ekaba Bisong. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners. New York, Apress, 2019, https://doi.org/10.1007/978-1-4842-4470-8.
22. Abadi, Martin, *et al.* “TensorFlow: A System for Large-Scale Machine Learning.” *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 2016, https://doi.org/10.48550/arXiv.1605.08695.