Using Data Augmentation to Improve the Performance of the ResNet-18 Model in Identifying Tuberculosis in Chest X-rays

Marco Sukhatme[1] and Parsa Akbari, Ph.D.[2]

[1]*Mamaroneck High School; msukhatme@gmail.com*
[2]*Department of Public Health and Primary Care, University of Cambridge; pa354@cam.ac.uk*

December 2022

1 Keywords

machine learning; convolutional neural networks; computational biology

2 Abstract

Tuberculosis is the second most lethal infectious disease in the world, but with early diagnosis and treatment, it is curable. However, with an examination cost of up to 175 U.S. dollars, many patients are not able to receive the diagnosis they need. X-ray imaging is used to identify patients who are at risk of tuberculosis and screen patients following completion of a treatment plan. Still, the assessment of X-ray images by a healthcare professional is time-consuming and expensive. This is particularly concerning considering the disease's prevalence in developing countries. It has been proposed that the diagnosis of tuberculosis from X-ray images can be automated with convolutional neural networks—models which rely on a training set to infer predictive features. However, these models do not generalize to new data external to the training set due to the variation in X-ray imaging techniques and quality which results in variations in orientation, brightness, contrast, saturation, noise, and resolution. In this study, a convolutional neural network is trained to diagnose tuberculosis from X-ray images and utilize augmentation to artificially insert variation into the training set images. The present study's methodology shows substantial improvements of up to 40% in model performance and six-fold reductions in training time.

3 Introduction

Tuberculosis (TB) is an infectious disease caused by the bacterium *Mycobacterium tuberculosis*. The bacterium most commonly infects the lungs through airborne human-to-human transmission, although it is also occasionally detectable in other organs such as the kidneys, brain, and spine. Following initial infection, the bacterium can further spread into the bloodstream and other organs. Typical symptoms of TB include fever, night sweats, and persistent coughing with blood-containing mucus, although most cases of TB remain latent showing no symptoms.

TB is still a prevalent disease in many parts of the world,[1] particularly in Africa and Southeast Asia, and it is often difficult to diagnose due to the cost of the examination of X-ray images by radiologists.[2] According to de Siqueira-Filha et al.,[3] the cost of diagnosis varies from $0.50 to $175 USD, primarily based on the complexity of the radiographic presentation. TB is often difficult to diagnose in patients with HIV infection,[4] which is common in the same areas that TB is common. Thus, undiagnosed patients are often forced to pay on the upper side of the aforementioned scale. However, machine learning models can automate the diagnosis of TB from chest X-ray images. Once diagnosed, patients would still need to be treated by a doctor, but automated diagnosis would significantly lower the price of treatment and make it more accessible.

One of the challenges with machine learning models is the need for a large dataset with reasonable variation, which is instrumental in training a good model. Moreover, X-rays from the United States are protected by the Health Insurance Portability and Accountability Act (HIPAA) and are considered private data, making them difficult to obtain. Due to the lack of data, there is often not enough variation to train the model to be as accurate as possible, since many of the images are uniform in their presentation. As a result, image data augmentations have been shown to be useful in improving variation within datasets.[5] By implementing these augmentations in training, the accuracy of the model could be improved.[6] The

present study used the ResNet-18 neural network[7] and multiple data augmentations—including variations in orientation, brightness, contrast, saturation, noise, and image resolution—to modify images within the dataset, and recorded how it affected the model's accuracy and ability to learn. All of the augmentations were expected to result in improved model performance.

4 Related Work

4.1 Data Augmentation

Previous studies have shown that data augmentation can lead to an improvement in the performance of a machine learning model. Gao et al.[6] found the use of data augmentations to be helpful in improving the performance of deep neural networks (DNNs). However, the present study aimed to instead test the improvement in the performance of a convolutional neural network (CNN), the difference between the two being that CNNs train slower in return for the reduction of the number of parameters.[8] Bhagoji et al.[5] also showed that data transformations can improve the performance of DNNs. These transformations are just a singular type of data augmentation though, specifically rotations, reflections, translations, scaling, etc. On the other hand, the present study combined the use of transformations—horizontal flip—with other types of data augmentation—variations in brightness, contrast, saturation, noise, and image resolution. Han et al.[9] did focus on the benefits of data augmentation on the performance of CNNs, but used the ILSVRC12 dataset. The present study used the ResNet-18 model, the Adam optimizer, and chest X-rays. In the study's analyses, known data augmentation techniques were applied in the context of training a convolutional neural network to predict TB diagnosis from X-ray images. In particular, augmentations that simulate natural variations normally observed in X-ray images and are likely to be encountered by a predictive model used in different clinical settings were selected.

4.2 ResNet-18

The ResNet-18 model has been shown to be effective in retrieving medical images.[10] Ayyachamy et al. showed the model's effectiveness in retrieving four specific types of medical images—computed tomography (CT), magnetic resonance imaging (MRI), mammogram (MG), and positron emission tomography (PET). However, this study did not test the ResNet-18 model's effectiveness in retrieving chest X-rays, which are the main imaging modality for TB globally. Shelke et al.[11] used the ResNet-18 model to retrieve chest X-rays, but the study attempted to identify COVID-19 in these images, whereas the present study attempted to identify TB. It is not clear whether the model trained by Shelke et al. would be robust to differences in X-ray image techniques that the model would encounter if employed as a predictive tool—for example, differences in brightness and saturation common amongst different X-ray imaging machines across different clinical sites.

5 Methods

5.1 Analysis Workflow and Types of Data Augmentation

Various types of data augmentation were used to test the model's ability to run efficiently under different circumstances, including variations in orientation, brightness, contrast, saturation, noise, and image resolution

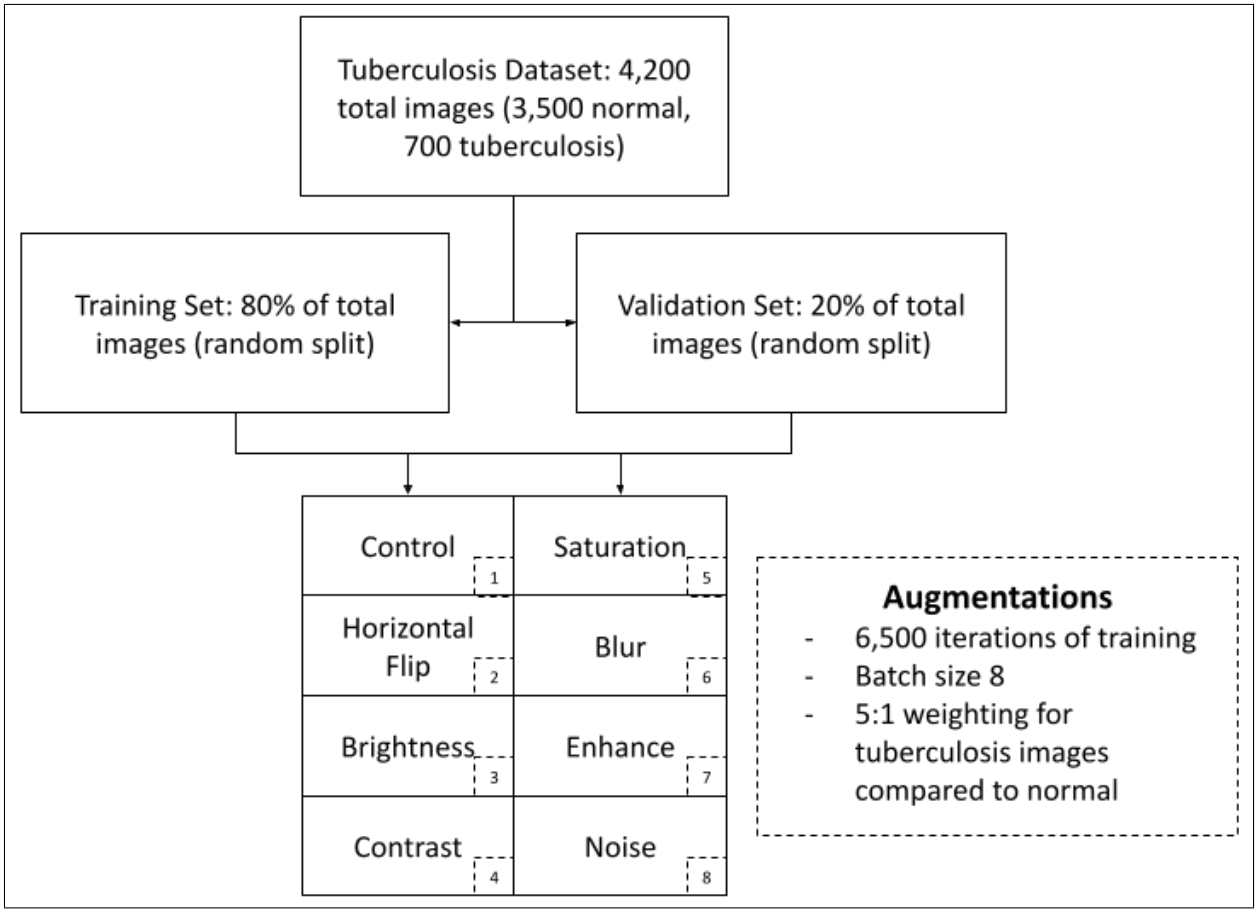

FIGURE 1. Analysis Workflow Including Dataset Split and Augmentations
An analysis workflow of the study, starting with the splitting of the dataset into training and validation sets, followed by the application of different data augmentations as shown. This is an overview of the experimental design for this study.

Experiment	Effect	Chance
Horizontal Flip	Mirroring across the y-axis	50%
Brightness	Random variation in brightness from 80% to 120%	100%
Contrast	Random variation in contrast from 80% to 120%	100%
Saturation	Random variation in saturation from 80% to 120%	100%
Noise	Random alteration of the intensity of each pixel	100%
Blur	Altered image being half as blurry as original image	50%
Enhancement	Altered image being twice as sharp as original image	50%

TABLE 1. Effect of Each Type of Data Augmentation
The effect of each type of data augmentation tested and the percent chance of it occurring.

(Table 1).

Different data augmentations are often most successful when they closely simulate new images that would naturally appear. Horizontal flip was chosen as one of the augmentations that the data would undergo because the lungs naturally appear symmetrical, making the augmented image seem relatively normal. Variations in brightness, contrast, saturation, noise, and image resolution were the other augmentations selected because they are some of the most common variables in chest X-rays, often varying depending on the quality and type of X-ray machine used to take the image.

Data augmentations were implemented in both the training and validation sets in order to test the model's improvement in accuracy. In addition, a control without any data augmentation in either of the sets was also run.

Data were randomly assigned to the training and validation sets, with 80% of images going into the training set. The model used was ResNet-18 and the optimizer was the Adam optimizer. A seed—a starting point for the sequence—was also implemented, resulting in a consistent sequence of numbers across multiple runs. The randomization of a machine learning algorithm is not actually random but in fact pseudorandom, so this seed adds consistency to the program.

Once the control experiment was determined to be sufficiently accurate by having a relatively low validation loss of 12.35% and a relatively high validation F-score of 94.91%, experiments were separately run with the different types of data augmentation—variations in orientation, brightness, contrast, saturation, noise, and image resolution.

5.2 Model and Optimizer

The increased depth of a neural network more easily allows the modeling of complex non-linear interactions between the input features, which is critical in achieving good performance in complex image recognition tasks. In this experiment, the ResNet-18 model[7] was used because its increased depth of 152 layers—eight times lower than that of VGG networks—combined with a lower complexity makes it very compatible with the identification of TB in chest X-rays. ResNet-18 has already been shown to be effective in retrieving medical images,[10] so the present study sought to determine its effectiveness in retrieving chest X-rays and

determine if it could be improved using data augmentations.

The Adam optimizer,[12] which is an improvement upon standard gradient descent in that it combines AdaGrad and RMSProp was used. AdaGrad provides the gradient descent with momentum so that it can get past large dips in the loss function, and RMSProp allows it to accelerate on flat planes. Default parameters were used for the Adam optimizer.

Each experiment was only run once, as a pseudorandom seed was used to make results consistent over any number of runs, and so multiple runs for each experiment would have been redundant.

5.3 Dataset

The dataset used included 4,200 chest X-ray images including 700 TB cases, sourced from Rahman et al.[13] Images were taken in the posterior-anterior orientation and images of TB cases showed the classical opacification of airspaces in the lung parenchyma and enlarged lymph nodes. Cavitary lesions were not observed, showing that the patients in the dataset were not advanced cases of TB.

5.4 Model Performance Metrics

The metrics used to measure the model's performance were precision, recall, F-score, and cross-entropy loss. Precision, recall, and loss measure the performance of the model (equations given below), while F-score is an advanced metric that combines both precision and recall. Cross-entropy loss measures the performance of a model that outputs a probability value between 0 and 1, increasing as the predicted probability diverges from the actual value.[14] True positive (TP) represents when the model correctly predicts when the image does indicate TB. False positive (FP) represents when the model incorrectly predicts when the image does indicate TB. True negative (TN) represents when the model correctly predicts when the image does not indicate TB. False negative (FN) represents when the model incorrectly predicts when the image does not indicate TB.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F-score} = \frac{2(\text{precision} + \text{recall})}{\text{precision} + \text{recall}} \tag{3}$$

6 Results

The present study trained the ResNet-18 convolutional neural network to predict TB from 4,200 images. Images were drawn randomly within each batch with weighting to ensure equal distribution of output classes, and the model was trained using an Adam optimizer with default parameters and a cross-entropy loss function.

Small training datasets are a major limiting factor in training effective machine learning models because they often lack variation. The use of data augmentations can effectively have the same result as a large data

Experiment	Validation F-score Percentage	Error Reduction Percentage
Horizontal Flip	97.10	43.03
Brightness	96.00	21.51
Blur	96.00	21.45
Saturation	95.20	5.77
Enhancement	94.95	0.81
Control	94.91	0.00
Contrast	94.33	-11.26
Noise	94.15	-14.83

TABLE 2. Validation F-score
The validation F-scores of experiments that have undergone different types of data augmentation at 6,500 training iterations.

set because they provide the additional variation that is needed to increase the model's performance. In order to assess the impact of data augmentations, variations in orientation, brightness, contrast, saturation, noise, and image resolution were selected.

6.1 ResNet-18 Is Able to Diagnose Tuberculosis With a Validation F-score of 95%

After 6,500 iterations of training, the validation F-score of the control experiment, using the ResNet-18 model, was 94.91% (Table 2, Figure 2). This is without any data augmentation, but could likely be improved with more iterations.

6.2 Horizontal Flip Had the Greatest Validation F-score Improvement Compared to Control of Any Augmentation Tested, at 97%

The horizontal flip experiment had the greatest validation F-score improvement compared to the control experiment of any data augmentation type tested. The horizontal flip experiment had a validation F-score of 97.10%, indicating a 2.19% improvement compared to the control experiment's validation F-score of 94.91% (Table 2, Figure 2). The experiment with the second greatest validation F-score improvement compared to the control experiment was the variations in brightness experiment, which had a validation F-score of 96.00%. This shows a 1.10% improvement compared to the control experiment (Table 2, Figure 2).

6.3 Augmentations Can Result in up to a 43% Reduction in Validation F-score Error Compared to Control

The validation F-score of the control experiment was 94.91% (Table 2, Figure 2), meaning that it is already relatively close to 100%. Further improvements in model performance become substantially more difficult as the F-score approaches 100%. Therefore, an increase from a validation F-score of 94.91% to 97.10% is

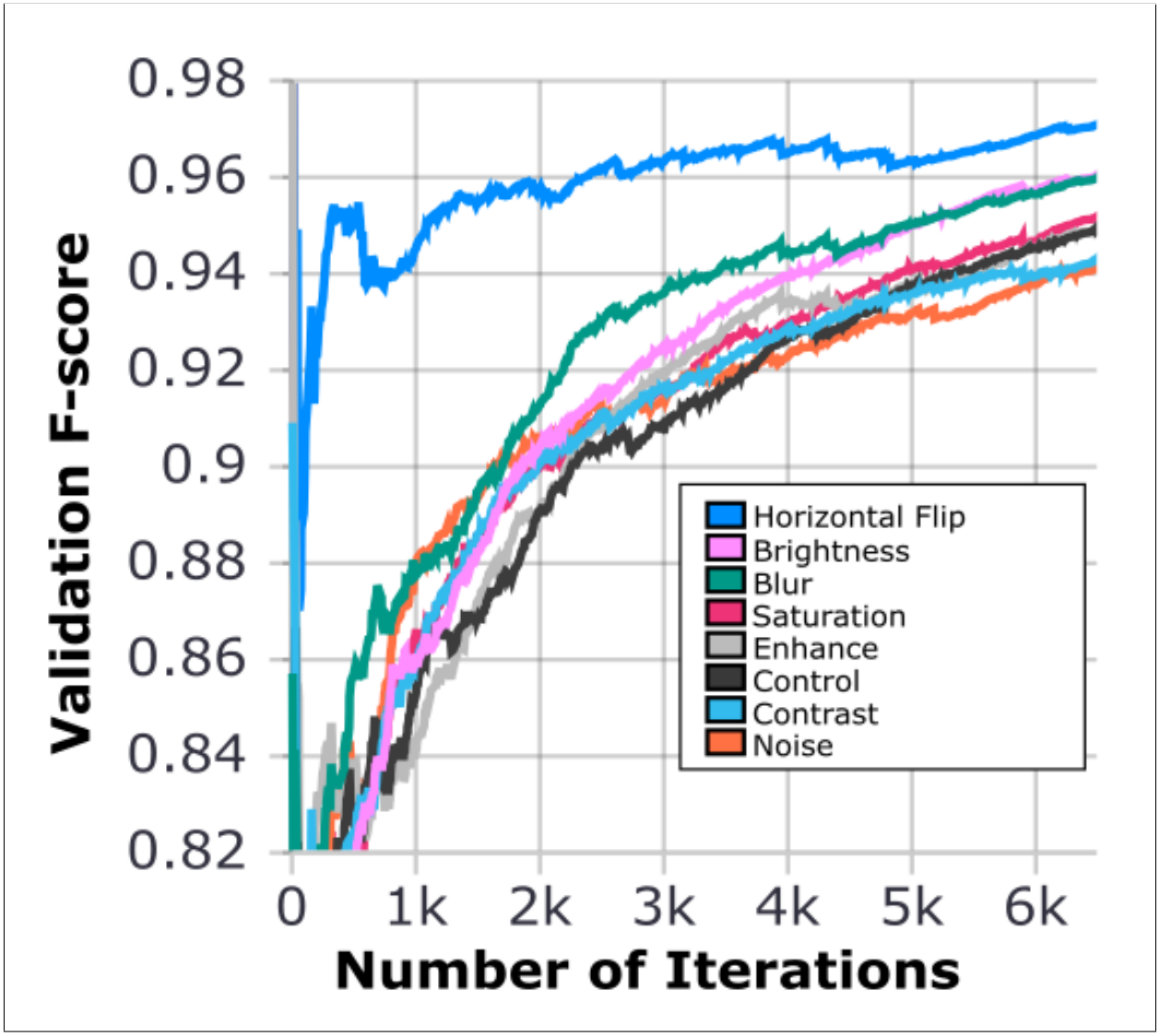

FIGURE 2. Validation F-score Training Curve Across Seven Augmentations
The validation F-score of experiments that have undergone different types of data augmentation over an amount of time, represented by the number of training iterations that the validation F-score was measured at.

significant in that it indicates a high reduction in error. This result was achieved using image augmentations to the dataset, which replicates the effect of a larger dataset in that it increases variation.

The validation F-score for the seven image augmentation experiments ranged from 94.15% to 97.10% with the highest being horizontal flip and the second highest being variations in brightness (Table 2, Figure 2). The validation F-score of the horizontal flip experiment was 97.10%, compared to the control experiment's validation F-score of 94.91%. Therefore, the horizontal flip experiment's validation F-score error was 2.90% and the control experiment's validation F-score was 5.09%. This is a reduction of 2.19%, or 43.03% of the original 5.09% error (Table 2, Figure 2).

The variations in brightness experiment had a validation F-score of 96.00%. Therefore, its validation F-score error was 4.00%, which, compared to the validation F-score error of the control experiment, is a reduction of 1.10%, or 21.51% of the control's 5.09% validation F-score error (Table 2, Figure 2).

6.4 Augmentations Can Result in up to a Six-Fold Reduction in Training Time Compared to Control

A leading factor in machine learning being inaccessible, specifically in healthcare, is the amount of time model training often takes. Machine learning workflows with long training times are expensive and difficult to run, resulting in a lack of interest in developing these models.

The control experiment reached its validation F-score of 94.91% after 6,500 iterations (Table 2, Figure 2). These 6,500 iterations took about 19 hours during the study, but the actual time it would take is dependent on the computer's performance.

The horizontal flip augmentation was able to achieve a validation F-score of greater than 94.91%, the control experiment's final validation F-score after 6,500 iterations, after only 309 iterations (Figure 2). This is equal to a 21.04 times training time reduction. However, the horizontal flip experiment's validation F-score dips below 94.91% at 558 iterations. It again rises above 94.91% and maintains this at 1,077 iterations (Figure 2). This indicates a 6.04 times training time reduction.

6.5 Contrast and Noise Found to Not Be Effective in Improving Validation F-score Compared to Control

It is common in model training to apply many different data augmentations to the dataset at one time, without fully understanding which augmentations most improve accuracy. This can lead to result deterioration from the overuse of these augmentations. In this analysis, seven data augmentations were tested and several augmentations that counterintuitively do not result in improvements in model performance despite the data augmentation reflecting variations that occur in real X-ray images were identified, including contrast and noise.

Contrast and noise are two data augmentations that are often found to be effective in improving a model's performance.[15] However, in the present study's analysis, these two augmentations were found to lead to a decrease in performance. The control experiment had a validation F-score of 94.9% after 6,500 iterations, while the contrast and noise experiments had validation F-scores of 94.3% and 94.2%, respectively (Table 2, Figure 2).

7 Discussion

The present study tested seven data augmentations that were chosen to reflect common variation in X-ray images. Augmentations have been shown to have the same result as increasing the size of the training set in that they improve dataset variation and model performance. The study identified image augmentations which resulted in a maximum of a 43% reduction in validation F-score error and a six-fold reduction in training time (Table 2, Figure 2). These results suggest that image augmentations, in particular horizontal flip and variations in brightness, are critical for model training on X-ray images. The initial hypothesis, that augmentations that reflect variations that are prevalent in X-ray images would improve the performance of the model, was rejected, as the augmentations for variations in contrast and noise did not result in an increase in validation F-score, instead resulting in a decrease.

Augmentations were initially selected under the criterion that the images would undergo on the basis of them frequently occurring in the real world. For example, X-ray images typically have differences in brightness and contrast, often based on the quality of the machine taking the image. This criterion may not necessarily be the best predictor for what would lead to improved model performance. Additionally, the degree of change to the image caused by each individual augmentation may have negatively affected the model's performance in the case of the augmentations for variation in contrast and noise. These augmentations may have caused the image to lose key identifying features, making it more difficult for the model to accurately diagnose them. The full extent of why some augmentations were not as successful as others would require further research to fully understand.

Further research could expand on the general scope of this study. One way that this study could be expanded would be by testing more types of data augmentation in addition to the seven that were tested. These augmentations could also be tested on various different models—the present study only tested on the ResNet-18 model. While the findings do imply that data augmentation can lead to performance improvement in all or at least most machine learning models, the study did not actually confirm this theory. The actual results, specifically the validation F-scores for all seven augmentation types and the control, could also be improved given more resources and time. A more powerful computer and/or multiple computers could greatly enhance the study in this aspect. In terms of understanding the findings, a study could be designed to determine which data augmentation types are not helpful in improving performance and why.

Reducing the cost of TB examination and diagnosis would allow patients to access the treatment they otherwise would not be able to. The present study trained a machine learning model to diagnose TB from chest X-ray images. Restrictions regarding patients' privacy results in a lack of access to a large dataset. This problem was circumvented by adding variation to the dataset in the form of image augmentations. The study identified augmentations that significantly improved model performance and reduced training time. The results from the study will inform the training of more accurate machine learning models to diagnose TB in chest X-rays, making treatment more reliable and cost-effective.

References

[1] Adam MacNeil, Philippe Glaziou, Charalambos Sismanidis, Anand Date, Susan Maloney, and Katherine Floyd. Global epidemiology of tuberculosis and progress toward meeting global targets - worldwide, 2018. *MMWR Morb. Mortal. Wkly. Rep.*, 69(11):281–285, March 2020.

[2] David Beal and Karen J Foli. Affordability in individuals' healthcare decision making: A concept analysis. *Nurs. Forum*, 56(1):188–193, January 2021.

[3] Noemia Teixeira de Siqueira-Filha, Rosa Legood, Aracele Cavalcanti, and Andreia Costa Santos. Cost of tuberculosis diagnosis and treatment in patients with hiv: a systematic literature review. *Value in health*, 21(4):482–490, 2018.

[4] Mbulelo Mntonintshi, Sikhumbuzo Mabunda, Kakia AF Namugenyi, and Don O'Mahony. Undiagnosed tuberculosis in patients with hiv infection who present with severe anaemia at a district hospital. *African Journal of Primary Health Care and Family Medicine*, 9(1):1–6, 2017.

[5] Arjun Nitin Bhagoji, Daniel Cullina, Chawin Sitawarin, and Prateek Mittal. Enhancing robustness of machine learning systems via data transformations. April 2017.

[6] Xiang Gao, Ripon K Saha, Mukul R Prasad, and Abhik Roychoudhury. Fuzz testing based data augmentation to improve robustness of deep neural networks. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, New York, NY, USA, June 2020. ACM.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. December 2015.

[8] Jiaxing Tan, Yumei Huo, Zhengrong Liang, and Lihong Li. A comparison study on the effect of false positive reduction in deep learning based detection for juxtapleural lung nodules: CNN VS DNN. In *Proceedings of the Symposium on Modeling and Simulation in Medicine*, pages 1–8. scs.org, 2017.

[9] Dongmei Han, Qigang Liu, and Weiguo Fan. A new image classification method using CNN transfer learning and web data augmentation. *Expert Syst. Appl.*, 95:43–56, April 2018.

[10] Swarnambiga Ayyachamy, Varghese Alex, Mahendra Khened, and Ganapathy Krishnamurthi. Medical image retrieval using resnet-18. In *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*, volume 10954, pages 233–241. SPIE, March 2019.

[11] Ankita Shelke, Madhura Inamdar, Vruddhi Shah, Amanshu Tiwari, Aafiya Hussain, Talha Chafekar, and Ninad Mehendale. Chest x-ray classification using deep learning for automated COVID-19 screening. *SN Comput Sci*, 2(4):300, May 2021.

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. December 2014.

[13] Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, Mohamed Arselene Ayari, and Muhammad E H Chowdhury. Reliable tuberculosis detection using chest X-Ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020.

[14] Thomas Phan. A brief tutorial on the relationship between maximum likelihood estimation and Cross-Entropy loss. 2021.

[15] Juan F Ramirez Rochac, Nian Zhang, Jiang Xiong, Jing Zhong, and Timothy Oladunni. Data augmentation for mixed spectral signatures coupled with convolutional neural networks. In *2019 9th International Conference on Information Science and Technology (ICIST)*, pages 402–407. ieeexplore.ieee.org, August 2019.