PAPER 69 – OPTIMIZING DEEP LEARNING FOR MEDICAL IMAGE CLASSIFICATION: A COMPARISON OF GRADIENT-BASED METHODS

Y. Ibrahim[1*], M. O. Momoh[2], K. O. Shobowale[3], Z. M. Abubakar[1], B. Yahaya[1], M. B. Mu'azu[1]

[1]Department of Computer Engineering, Ahmadu Bello University, Zaria, Nigeria.
[2]Department of Aerospace Engineering, Air Force Institute of Technology, Kaduna, Nigeria.
[3]Department of Mechatronics Engineering, Air Force Institute of Technology, Kaduna, Nigeria.
[*]Email: yibrahim@abu.edu.ng

ABSTRACT
Deep learning-based medical image classification has shown promising results in recent years with performance heavily reliant on choice of optimization schemes during training. In this paper, we present a comparative study of seven first-order stochastic gradient-based optimization schemes for a medical image classification task. A custom CNN architecture with convolutional layers and increasing number of filters followed by max-pooling and ReLU activation function with Dropout regularization were trained on a Tuberculosis Chest X-ray Database. The models were tested on a held-out test set to assess their performance in terms of accuracy, precision, recall, F1-score, and AUC score. Our experimental results demonstrate that Adamax achieves the highest accuracy and F1-score of 99.00% each and highest AUC score of 99.80%. Adadelta and Adagrad performed relatively poorly in comparison to other optimizers, achieving accuracies of 82.14% and 88.00% respectively. In terms of recall, both Adam and Nadam gave the highest recall scores of 98.86%. This suggests that the choice of optimizer plays a significant role in the performance of these models and provides valuable insights into the suitability of different optimizers for medical image classification. This will be beneficial for researchers and practitioners in the field of medical image analysis using deep learning.

KEYWORDS: Deep Learning, Convolutional Neural Networks, Optimizers, Medical Image Classification, Chest X-Ray Classification

1. INTRODUCTION

Medical image classification has nowadays become an increasingly important area of research, driven by advances in deep learning (DL) techniques and the growing availability of large-scale medical image datasets (Çallı *et al*., 2021). It is a critical task in healthcare and can assist professionals such as radiologists in detecting diseases and anomalies more accurately and efficiently. Among the various medical imaging modalities, X-ray is one of the most widely used *(Çallı et al*., 2021) due to its low cost and non-invasive nature with the ability to provide fast results. In particular, it has been used to detect Tuberculosis (TB), a chronic respiratory illness caused by bacterial infection which poses significant global health problem. According to the World Health Organization (WHO), TB is the 13th leading cause of death worldwide and the second leading infectious killer after COVID-19 (above HIV/AIDS)(Organization, 2022). In 2021, a total of 1.6 million people died from TB(Organization, 2022). Timely detection and accurate identification of TB are hence vital as delaying the diagnosis could be fatal, and prompt administration of appropriate medication can cure this lethal ailment (Sharma & Mohan, 2013).Therefore, there is need to develop schemes for accurate and timely diagnosis for effective treatment and control of these diseases, especially in low-resource settings where they are prevalent.

The traditional machine learning (ML) methods for medical image analysis typically involve hand-crafted extraction of features and feeding them to classical classification algorithms such as random forests (RF) (Melendez *et al.*, 2016), extremely randomized trees (ERT)(Melendez *et al.*, 2016), k-nearest neighbors (KNN) (Van Ginneken *et al*., 2002), etc. This approach relies on prior knowledge of the image content and may not generalize well to new datasets or imaging modalities (Melendez *et al*., 2016; Singh & Hamde, 2019; Van Ginneken *et al.*, 2002). These methods also require extensive manual annotation and may be prone to errors (Rahman *et al*., 2020), variability and subjective interpretation (Brady, 2017; Degnan *et al.*, 2019).

In contrast, the availability of extensive labeled datasets and deep convolutional neural networks (CNNs) has resulted in significant achievements in image recognition. CNNs can learn data-driven, highly descriptive, and hierarchical image features from a sufficient amount of training data. However, it is still challenging to acquire comprehensive annotations for medical imaging datasets like ImageNet (Greenspan *et al*., 2016; Shin *et al*., 2016). Furthermore, continued research in this area has led to significant advances in the diagnosis and treatment of a wide range of diseases.

Also, the effectiveness of these models is heavily reliant on the choice of optimization algorithms used during training. Gradient-based optimization algorithms have been widely adopted due to their efficiency and effectiveness in DL tasks (Dogo *et al*., 2018). However, selecting the optimal optimization scheme for a specific

task still remains a challenging and important problem especially for medical image detection and classification tasks.

This paper therefore presents a comparative study of popular first-order stochastic gradient-based optimization schemes for medical image classification, specifically focusing on optimizing DL models for TB Chest X-ray (CXR) Database. We analyze the performance of different optimizers in terms of accuracy, precision, recall, F1-score, and AUC score and provide insights into their suitability for medical image classification.

This study contributes to the growing body of literature on the utilization of DL models for medical image detection and classification. The findings of this study will be of great interest to researchers and practitioners in the field of medical image analysis using DL, as it provides valuable insights into the suitability of different optimizers for medical image classification. We evaluate our trained custom CNN models on a held-out test data and analyze their performance using various evaluation metrics and visualization techniques with a view to addressing the following research questions:

- How does the choice of gradient-based optimization schemes affect the performance of DL-based medical image classification models?
- Which first-order stochastic gradient-based optimization scheme performs the best on a custom CNN architecture for the TB CXR image classification in terms of accuracy, precision, recall, F1-score, and AUC score?
- What insights can be gained from comparing the performance of the various optimization schemes and how can this information be used to guide the selection of optimizers for medical image classification tasks?

The objectives of the study include providing an empirical evaluation of the impact of popular gradient-based optimization schemes on the performance of DL-based medical image classification models, designing and implementing a custom CNN architecture and comparing the performance of various first-order stochastic gradient-based optimization schemes. The paper also offers a benchmark for future researchers and practitioners to compare their results and select the most appropriate optimizer for their medical image classification task and provide guidance for the selection of optimization schemes for medical image classification tasks.

2. RELATED WORK

In various recent research, CNNs have been utilized to identify a number of lung ailments like TB and pneumonia through the analysis of chest X-ray (CXR) images. With the emergence of the COVID-19 pandemic in 2020, techniques based on CNNs have also been employed to detect the presence of the novel coronavirus by examining CXR images. The work of Sathitratanacheewin *et al*., (2020) highlights the limitations of using ML models for TB detection and surveillance, specifically with regards to the generalizability of the models. The study developed a Deep CNN model using a TB-specific CXR dataset from one population and tested it on a non-TB-specific CXR dataset from another population. The CCN model performed well in detecting TB in both the training and intramural test sets using the former, with AUC values of 0.9845 and 0.8502, respectively. However, the AUC dropped significantly to 0.7054 in the latter showing that a supervised DL model developed using the training dataset from one population may not have the same diagnostic performance in another population, indicating the importance of examining technical specifications of CXR images, disease severity distribution, dataset distribution shift, and overdiagnosis before implementing the model in other settings. The work of Rahman *et al*., (2020) developed a framework using CXR with DL, segmentation and visualization. The work was aimed to detect TB reliably from CXR images by utilizing image preprocessing, data augmentation, image segmentation, and deep-learning classification techniques. The study created a database of 3500 TB-infected and 3500 normal CXR images and utilized nine different deep CNNs for transfer learning. The study found that classification using segmented lung images outperformed that with whole X-ray images and achieved state-of-the-art performance with accuracy, precision, sensitivity, F1-score, and specificity of DenseNet201 being 98.6%, 98.57%, 98.56%, 98.56%, and 98.54% respectively. The proposed method could be useful in computer-aided faster diagnosis of TB. In their study, Tahir *et al*., (2021) achieved a sensitivity of over 90% in classifying several coronavirus families including SARS, MERS, and COVID-19 through transfer learning of various pre-trained CNN models. Chhikara *et al*., (2020) investigated whether CXR images could be used to identify cases of pneumonia. They assessed the effectiveness of several pre-trained models, including Resnet, Xception, and Inception, while also implementing various preprocessing techniques such as filtering and gamma correction. In several studies, deep ML algorithms have been utilized to detect TB by adjusting the parameters of deep-layered CNNs (Hooda *et al*., 2017; Lakhani & Sundaram, 2017; Nguyen *et al*., 2019; Pasa *et al*., 2019). Pre-trained models and their ensembles were utilized for TB detection using transfer learning in the DL framework, as explained in sources (Ahsan *et al*., 2019; Meraj *et al*., 2019). A DL approach was presented in (Hooda *et al*., 2017) to classify CXR images into TB and non-TB categories achieving an accuracy of 82.09% while Evalgelista & Guedes, (2018) reported an accuracy of 88.76% for an intelligent pattern recognition framework based on CNNs for TB detection from CXR images. The work of Pasa *et al*., (2019) suggested a deep neural network design that was enhanced for detecting TB, achieving an accuracy of 86.82%. Additionally, they presented a tool that allows for the interactive visualization

of cases of TB. In their study, Nguyen *et al*., (2019) assessed how well the pre-trained DenseNet model could classify images of normal and TB cases from two different databases. They used a fine-tuned version of the model and found that it achieved AUC scores of 0.94 and 0.82. In their research, Ahsan *et al*., (2019) put forth a CNN model that was pre-trained in a general manner to detect TB and achieved accuracies of 81.25% and 80% with and without the use of image augmentation techniques, respectively. Abbas *et al*., (2020) proposed a CNN design that applies a class decomposition method to enhance the efficiency of ImageNet pre-trained CNN models for transfer learning. This approach resulted in excellent accuracy for detecting TB on the Japanese Society of Radiological Technology (JSRT) database. Yadav *et al*., (2018) utilized transfer learning to identify TB and achieved a precision rate of 94.89%.

3. MATERIALS AND METHODS

3.1. Dataset description

We used the TB Chest X-ray Database made publicly available by Rahman *et al*., (2020) containing 7,000 CXR images (3500 TB, 3500 normal) labelled as either TB (positive) or Normal (negative). For our experiments, we utilized the dataset in an 80:10:10 split for the training, validation, and testing sets respectively. As a preprocessing step, we resized the images to 224x224 and normalized by scaling the pixel values to [0,1].

3.2 Architecture

We trained a custom CNN architecture consisting of four convolutional layers with increasing number of filters (32, 64, 128, 256) followed by max pooling layers of size 2x2, which reduces the spatial dimensions of the feature maps. The ReLU activation function is used for all convolutional layers to introduce non-linearity into the model. A flatten layer is added to convert the 2D feature maps into a 1D vector, which is then passed through two fully connected dense layers with 512 and 1 neurons respectively. The first dense layer uses the ReLU activation function while the last layer uses the sigmoid activation function, which outputs a probability value between 0 and 1 for the binary classification task. Dropout regularization with a rate of 0.5 was used in the first dense layer to prevent overfitting. The input shape of the model is defined as (224, 224, 3) to accommodate 3 colour channels. A summary of the network configuration is shown in Table 1.

Table1: Summary of the network configuration

Dataset: TB Chest X-Ray Database
Input image data - 224x224x3- channels RGB Images
Conv3x3-32; stride=1
ReLU (nonlinearity function)
Pooling layer: MaxPooling 2x2; stride= 1
Conv3x3-64; stride=1
ReLU (nonlinearity function)
Pooling layer: MaxPooling 2x2; stride= 1
Conv3x3-128; stride=1
ReLU (nonlinearity function)
Pooling layer: MaxPooling 2x2; stride= 1
Conv3x3-256; stride=1
ReLU (nonlinearity function)
Pooling layer: MaxPooling 2x2; stride= 1
Flattening (Input Layer of Neural Network)
FC-512 neurons
ReLU (nonlinearity function)
Dropout = 0.5 (regularization to prevent overfitting)
Final dense layer
Binary Cross entropy
Sigmoid Layer (ϕ)

3.3 First-order stochastic gradient-based optimization schemes

We explored a class of optimization algorithms that use the gradient of the loss function with respect to the model parameters to update the parameters iteratively during training. These methods differ in their approach to adaptively adjusting the learning rate, incorporating momentum, or scaling the gradient updates based on the history of the gradients or the model parameters. For the custom CNN architecture in Table 1, we experimented with seven gradient-based optimizers including Adaptive Moment Estimation (Adam), Adaptive Moment Estimation with an infinity norm extension (Adamax), Nesterov-accelerated Adaptive Moment Estimation (Nadam), Root Mean Square Propagation (RMSProp), Stochastic Gradient Descent (SGD), Adaptive Delta (AdaDelta), and Adaptive Gradient (AdaGrad). The learning rate was set as 0.001 with a batch size of 32 and each model was trained for 20 epochs. All other training parameters were set at their default settings.

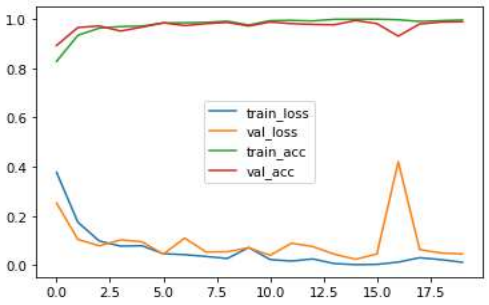
3.4 Evaluation metrics

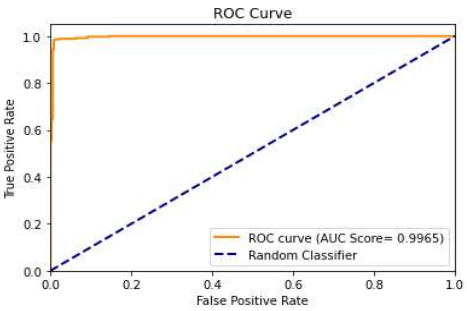
We evaluated the performance of the models using several standard evaluation metrics, including accuracy, precision, recall, and F1-score. The accuracy measures the overall classification performance, while precision and recall measure the performance of the model for the TB (positive) and Normal (negative) classes, respectively. F1-score is the harmonic mean of precision and recall and provides an overall measure of the model's performance. We also generated the Receiver Operating Characteristic (ROC) to visualize the model's performance and noted the Area under the curve (AUC) performance which measures the performance of the binary classifier.

4. RESULTS AND DISCUSSION

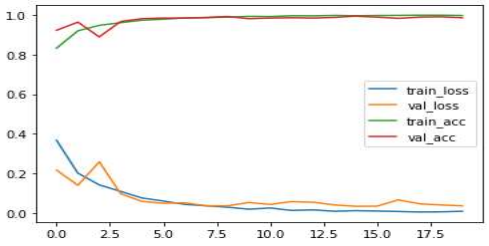
In this study, we trained and evaluated seven models (Models 1-7) corresponding to model trained with Adam, Adamax, Nadam, RMSprop, SGD, Adadelta, and Adagrad) on the TB Chest X-ray Database of 7,000 images of TB and Normal classes. All models were based on the CNN architecture described in Table I. The ReLU activation function is used for all convolutional layers to introduce non-linearity into the model and Dropout regularization with a rate of 0.5 was used in the first dense layer to prevent overfitting. The training and validation loss and accuracy for all the seven models are as shown in Figure 1.

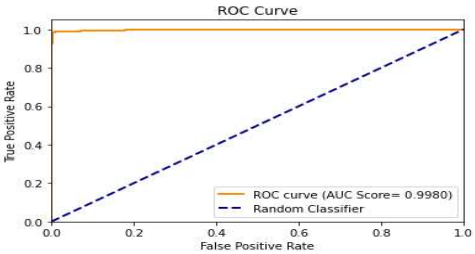
Adam



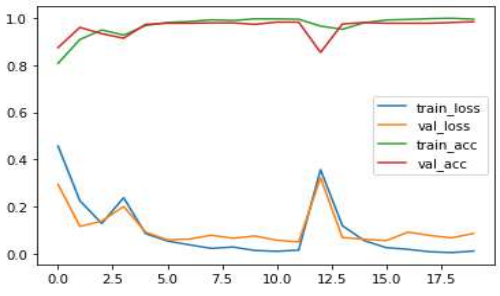


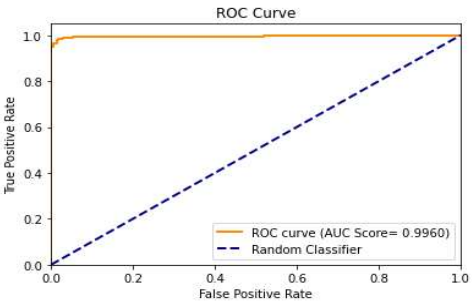
Adamax



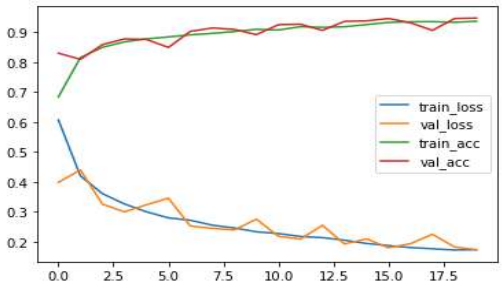


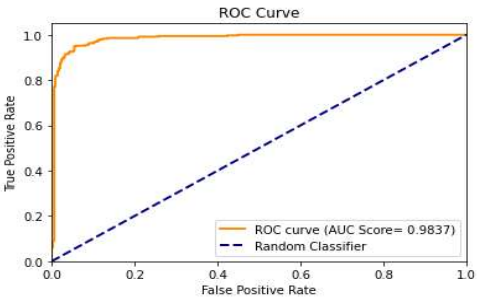
Nadam





Sgd





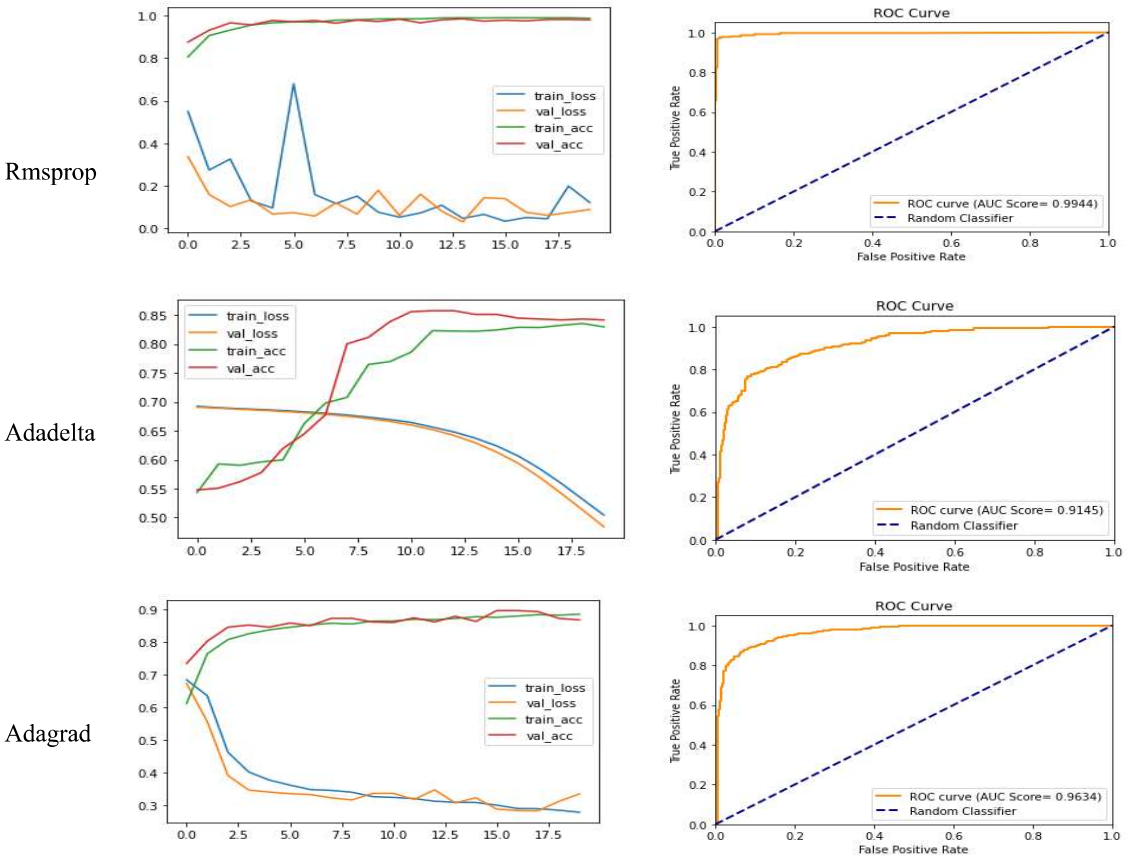


Figure 1: Training, validation, and testing performance of the various optimizers

Table 2 summarizes the performance of each optimizer in terms of the stated performance measures.

Table 2: Performance reports of various optimizers

Model	Optimizer	Accuracy	Precision	Recall	F1-Score	AUC score
Model 1	Adam	0.9800	0.9719	0.9886	0.9802	0.9965
Model 2	Adamax	0.9900	0.9942	0.9857	0.9900	0.9980
Model 3	Nadam	0.9757	0.9638	0.9886	0.9760	0.9960
Model 4	RMSprop	0.9786	0.9799	0.9771	0.9785	0.9944
Model 5	SGD	0.9414	0.9668	0.9143	0.9398	0.9837
Model 6	Adadelta	0.8214	0.7892	0.8771	0.8309	0.9145
Model 7	Adagrad	0.8800	0.9683	0.7857	0.8675	0.9634

Figure 2 shows the performance comparison of different optimizers based on several evaluation metrics (Accuracy, Precision, Recall, F1-Score and AUC Score.)

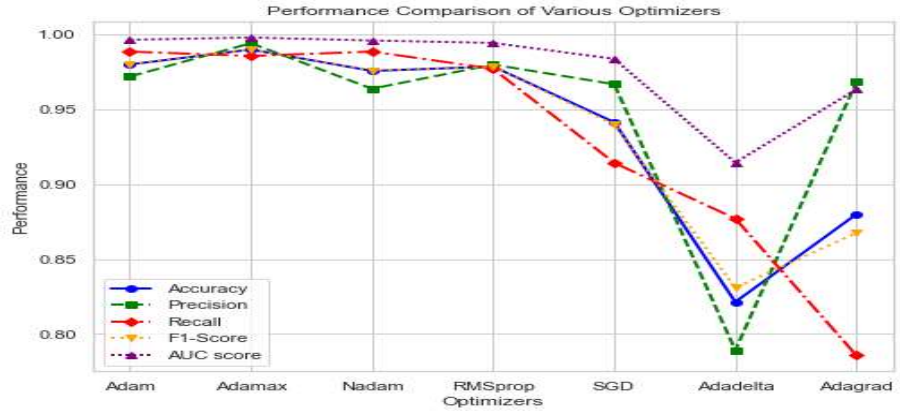


Figure 2: Comparison of optimizer performace

From the Figure 2, we can see that different optimizers perform differently for each evaluation metric. For example, Adamax and Adam appear to be the best optimizers for accuracy, while Adadelta gave the worst performance for precision, accuracy and f1-score. Specifically, the Adamax optimizer had the highest accuracy (99.00%) and performed best overall, with the highest precision (99.42%), F1-score (99.00%), and AUC score (0.9980) among all the tested optimizers. The Adam optimizer had the second-best accuracy (98.00%) and performed well, with a high precision (97.19%), recall (98.86%), and F1-score (98.02%). The RMSprop and Nadam optimizers also performed relatively well, achieving high accuracy, precision, recall, and AUC scores. However, the Adagrad and Adadelta optimizers were less effective, with lower accuracy (88% and 82.14%), F1-score (86.75% and 83.09%), and AUC scores (96.34% and 91.45%) respectively. The Adamax optimizer was the most effective in this study, achieving the highest scores across all metrics. The performance of our best-performing model (Model 2) is comparable to that of the state-of-the-art performance of Rahman *et al*., (2020) on the same dataset which achieved best accuracy of 98.6% despite utilizing DenseNet201 using transfer learning in classifying TB and normal images. We confirm that a less complex CNN architecture with proper parameter optimization can perform competitively with more complex state-of-the-art schemes.

5. CONCLUSION

In this study, we evaluated the performance of several popular first-order stochastic gradient-based optimization algorithms for optimizing a custom deep CNN weight for chest x-ray image classification. Our results demonstrated that a good choice of optimizer has a significant impact on the performance of DL models for CXR image classification with the best model achieving the highest testing accuracy and F1-Score of 99% each using Adamax optimizer. This is closely followed by Adam which recorded an accuracy and F1-Score of 98% and 98.02% respectively. In contrast, Adadelta and Adagrad optimizers resulted in comparatively poor performance, indicating that they may not be well-suited for this task. Interestingly, the AUC scores range from 0.9145 to 0.9980, indicating that the models built using these optimizers can distinguish between the positive and negative cases with varying degrees of accuracy. The highest AUC score achieved by the Adamax optimizer suggests that it is the most effective optimizer for this particular task. The study also suggests that the choice of optimizer should be based on the specific requirements of the task, and a thorough comparison of multiple optimizers should be conducted before making a final decision. The findings of this study provide useful insights for researchers and practitioners working on DL-based medical image analysis tasks. It would be of interest to investigate the effects of other CNN architectural designs and utilizing various medical datasets such as MRIs or CT scans for better generalization. We leave this exploration for future research. Other research directions could be developing explainability methods to aid physicians in interpreting the models' predictions and examining the models' impact on clinical outcomes in real-world settings to assess their efficacy in clinical practice.

REFERENCES

Abbas, A., Abdelsamea, M. M., & Gaber, M. M. (2020). Detrac: Transfer learning of class decomposed medical images in convolutional neural networks. *IEEE Access*, *8*, 74901-74913.

Ahsan, M., Gomes, R., & Denton, A. (2019). Application of a convolutional neural network using transfer learning for tuberculosis detection. 2019 IEEE International Conference on Electro Information Technology (EIT),

Brady, A. P. (2017). Error and discrepancy in radiology: inevitable or avoidable? *Insights into imaging*, *8*, 171-182.

Çallı, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G., & Murphy, K. (2021). Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis*, *72*, 102125.

Chhikara, P., Singh, P., Gupta, P., & Bhatia, T. (2020). Deep convolutional neural network with transfer learning for detecting pneumonia on chest X-rays. Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals: Proceedings of GUCON 2019,

Degnan, A. J., Ghobadi, E. H., Hardy, P., Krupinski, E., Scali, E. P., Stratchko, L., Ulano, A., Walker, E., Wasnik, A. P., & Auffermann, W. F. (2019). Perceptual and interpretive error in diagnostic radiology—causes and potential solutions. *Academic radiology*, *26*(6), 833-845.

Dogo, E. M., Afolabi, O., Nwulu, N., Twala, B., & Aigbavboa, C. (2018). A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks. 2018 international conference on computational techniques, electronics and mechanical systems (CTEMS),

Evalgelista, L. G. C., & Guedes, E. B. (2018). Computer-aided tuberculosis detection from chest X-ray images with convolutional neural networks. Anais do XV Encontro Nacional de Inteligência Artificial e Computacional,

Greenspan, H., Van Ginneken, B., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE transactions on medical imaging*, *35*(5), 1153-1159.

Hooda, R., Sofat, S., Kaur, S., Mittal, A., & Meriaudeau, F. (2017). Deep-learning: A potential method for tuberculosis detection using chest radiography. 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA),

Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, *284*(2), 574-582.

Melendez, J., Sánchez, C. I., Philipsen, R. H., Maduskar, P., Dawson, R., Theron, G., Dheda, K., & Van Ginneken, B. (2016). An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. *Scientific reports*, *6*(1), 25265.

Meraj, S. S., Yaakob, R., Azman, A., Rum, S., Shahrel, A., Nazri, A., & Zakaria, N. F. (2019). Detection of pulmonary tuberculosis manifestation in chest X-rays using different convolutional neural network (CNN) models. *Int. J. Eng. Adv. Technol.(IJEAT)*, *9*(1), 2270-2275.

Nguyen, Q. H., Nguyen, B. P., Dao, S. D., Unnikrishnan, B., Dhingra, R., Ravichandran, S. R., Satpathy, S., Raja, P. N., & Chua, M. C. (2019). Deep learning models for tuberculosis detection from chest X-ray images. 2019 26th international conference on telecommunications (ICT),

Organization, W. H. (2022). *Tuberculosis Fact Sheet*. Retrieved 27 October 2022 from https://www.who.int/news-room/fact-sheets/detail/tuberculosis

Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., & Pfeiffer, D. (2019). Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization. *Scientific reports*, *9*(1), 6268.

Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., Hamid, T., Islam, M. T., Kashem, S., & Mahbub, Z. B. (2020). Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access*, *8*, 191586-191601.

Sathitratanacheewin, S., Sunanta, P., & Pongpirul, K. (2020). Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability. *Heliyon*, *6*(8), e04614.

Sharma, S. K., & Mohan, A. (2013). Tuberculosis: From an incurable scourge to a curable disease-journey over a millennium. *The Indian journal of medical research*, *137*(3), 455.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, *35*(5), 1285-1298.

Singh, N., & Hamde, S. (2019). Tuberculosis detection using shape and texture features of chest X-rays. Innovations in Electronics and Communication Engineering: Proceedings of the 7th ICIECE 2018,

Tahir, A., Qiblawey, Y., Khandakar, A., Rahman, T., Khurshid, U., Musharavati, F., Islam, M., Kiranyaz, S., & Chowdhury, M. (2021). Coronavirus: Comparing COVID-19, SARS and MERS in the eyes of AI.

Van Ginneken, B., Katsuragawa, S., ter Haar Romeny, B. M., Doi, K., & Viergever, M. A. (2002). Automatic detection of abnormalities in chest radiographs using local texture analysis. *IEEE transactions on medical imaging*, *21*(2), 139-149.

Yadav, O., Passi, K., & Jain, C. K. (2018). Using deep learning to classify X-ray images of potential tuberculosis patients. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),