LIME-RAY: What Does a Neural Network See from X-Rays?

Abstract—**In the field of medical imaging, neural networks have significantly enhanced the analysis and interpretation of X-ray images, providing advanced capabilities in detecting patterns and anomalies. Despite their potential, challenges remain, particularly concerning the interpretability and reliability of these models. The "LIME-RAY" approach addresses these challenges by integrating Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks. This hybrid technique leverages CNNs for spatial feature extraction and LSTMs for temporal pattern recognition, improving the accuracy and interpretability of the model's predictions. Furthermore, Local Interpretable Model-agnostic Explanations (LIME) are used to provide transparent, interpretable results, crucial for gaining the trust of healthcare professionals. Our approach is validated using a diverse dataset of chest X-rays, including cases of Covid-19, Pneumonia, and Tuberculosis. The proposed method demonstrates superior performance in multi-class disease classification and offers valuable insights into the decision-making process, bridging the gap between complex neural network models and practical healthcare applications.**

Index Terms—**Chest X-ray, CNN-LSTM, Explainable AI, LIME, Pneumonia, Tuberculosis, healthcare applications**

I. Introduction

Neural networks have revolutionized the field of medical imaging, particularly in the analysis and interpretation of X-ray images. These advanced algorithms can detect patterns and anomalies with accuracy and speed that surpass traditional methods, thereby aiding radiologists in diagnosing conditions more efficiently [1], [2]. Neural networks, using vast datasets, enhance early disease detection and treatment outcomes by identifying subtle features in X-ray images, making them crucial in the medical sector for precise diagnostics [3], [4]. Despite their potential, neural networks in X-ray imaging are not without challenges [5]. One significant issue is the interpretability of the models; it is often unclear how these networks arrive at their conclusions. This can be problematic in clinical settings where understanding the rationale behind a diagnosis is crucial. This lack of transparency can lead to a lack of trust among healthcare professionals, who may be hesitant to rely on a "black box" system [6]. Additionally, neural networks can be prone to overfitting, especially when trained on limited or imbalanced datasets, leading to less reliable performance on new, unseen data [7]. Variability in image quality and differences in imaging protocols across institutions further complicate the application of these models in practice, potentially resulting in inconsistent diagnostic outcomes [8].

The authors propose a hybrid technique combining Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks to improve interpretability and robustness in medical diagnostics. The technique, called LIME-Ray, uses CNNs for spatial feature extraction and LSTMs for temporal pattern recognition, providing clear visual and analytical insights into the decision-making process. This approach bridges the gap between advanced neural network capabilities and practical healthcare needs, enhancing confidence and accuracy in medical diagnostics. Our research is centred around three key goals:

1) **Dataset Diversity**: We are working with a large and diverse dataset that includes X-ray images of three distinct types of respiratory diseases: Covid-19, Pneumonia, and Tuberculosis. This extensive dataset allows for a comprehensive analysis and ensures the robustness of our model across different conditions.
2) **Innovative Technique**: We employ a unique CNN+LSTM technique that combines the spatial feature extraction capabilities of CNNs with the temporal sequence learning of LSTMs, enhancing the model's ability to analyze complex medical images.
3) **Interpretability with LIME**: We introduce LIME in our multiclass classification approach to provide transparent and interpretable results, thereby gaining the trust of doctors and other healthcare professionals who rely on these systems for accurate diagnostics.

II. Literature Review

Chest X-ray (CXR) images and CT-scan image-based illness identification have been significantly improved by utilizing machine learning and deep learning models in medical image analysis. Transfer Learning (TL) has also gained acceptance for complex classification tasks in medical imaging, particularly in detecting disorders like COVID-19, lung infections, and liver cirrhosis. [9].

ResNet, GoogLeNet, Visual Geometry Group Network (VGGNet), and MobileNet are among the main classification methods for CXR images. All of these methods employ the TL technique inside the DL framework. Alqudah et al. and Oyelade et al. utilized pre-trained TL to diagnose pneumonia [10]. They constructed a computer-assisted diagnostic tool that leverages DL models to precisely identify pneumonia by analysing CXR images. Mehmood et al. utilized TL in the AlexNet architecture to boost the accuracy and efficiency of imaging histological lung and colon tissues [11]. Sitaula et al. concluded that CXR picture resolution varies in real-world applications, and a single-scale bag of deep visual words BoDVW-based features are insufficient to capture the semantic information of infected lung areas [12]. DL model with nine hidden layers was proposed by Mahbub et al. for CXR image categorization [13]. The DL model was used to classify COVID-19, TB, and Pneumonia using six publicly accessible CXR images. Despite its ability to break down disease classes, the model was tested on six datasets with two classes each and without explanation. Likewise, Qaqos et al. [14] trained a DL model with 6587 CXR images using SGD. Applying 128x128 images and 100 epochs, the model sorted CXR pictures into four categories (COVID-19, Pneumonia, and TB) with 94.53% accuracy. However, no XAI algorithms were implemented. In addition to TL, Sitaula et al. [15] adopted an attention mechanism alongside VGG16 to classify CXR pictures. The VGG16 and attention model were used to extract a particular region from CXR pictures, which were then depicted using Grad-CAM [16]. Ahsan et al. examined 400 CXR and 400 CT scans to detect COVID-19 patients using different versions of MobileNet, VGG and ResNets. Later, they applied the class activation heatmap to the predictions [17].

Our findings suggested that CXR is ideal for detecting diseases such as COVID-19, pneumonia, and tuberculosis. A singular model with prediction explaining ability can aid medical practitioners in diagnosis while gaining their trust.

III. Methodology

A. Dataset

The dataset used for this research was collected from several sources [18]–[20]. A total of 7,135 chest X-ray images consisting of 4 different classes. The images were divided into train, test, and validation sets. Each category contains sub-folders with images classified into four conditions: Tuberculosis (TB), COVID-19, and Pneumonia and Normal. The dataset, however, suffers from class imbalance.

B. Deep Learning Approach

As this is a multiclass image classification problem, we opted for deep learning instead of traditional machine learning. Several models were tested in the same setup. The proposed architecture of the CNN-LSTM hybrid model is described separately in the next section. Figure 1 depicts the total workflow of the LIME-Ray process.

1) AlexNet: AlexNet, a groundbreaking convolutional neural network (CNN) architecture, significantly influenced deep learning and computer vision [21]. It comprises 8 layers, employing techniques like ReLU activation, LRN, random cropping, and colour jittering. AlexNet demonstrated the superiority of CNNs over conventional image categorization techniques.

2) ResNet101 V2: ResNet101v2 is a highly regarded deep learning architecture, particularly effective in image classification tasks [22]. Its residual blocks, incorporating identity mappings, use a bottleneck design to reduce computational costs and employ pre-activation residual units to enhance the training process.

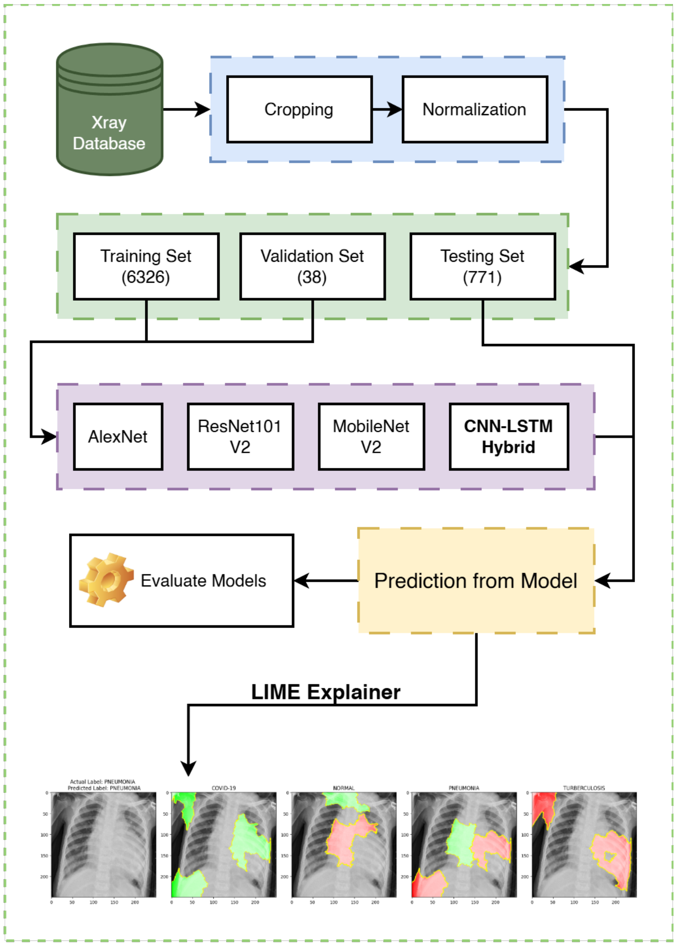

Fig. 1. Total workflow of LIME-Ray approach

3) MobileNet V2: MobileNet V2 is a well-designed architecture for mobile and on-board vision tasks, achieving a balance between computational speed and accuracy in object recognition [23]. Its compatibility with embedded platforms and depth-wise separable convolutions reduce computational costs.

C. Local Interpretable Model-agnostic Explanations (LIME)

The Local Interpretable Model-Agnostic Explanation (LIME) is a framework that provides localized explanations for black-box classification models [24]. It identifies the smallest features contributing to the highest probability of a class outcome for a single observation. This study aims to assess the efficacy of LIME explanations and their impact on healthcare professionals' trust in black-box classification algorithms. LIME has multiple stages, including selecting an occurrence, modifying data, forecasting results, and training a transparent model.

IV. Proposed CNN-LSTM Architecture

Fig. 2, depicts the skeletal layout of the proposed CNN-LSTM hybrid model. Firstly, A batch of 16 images, each image of 250x250x3, is taken as input. Then, the model uses

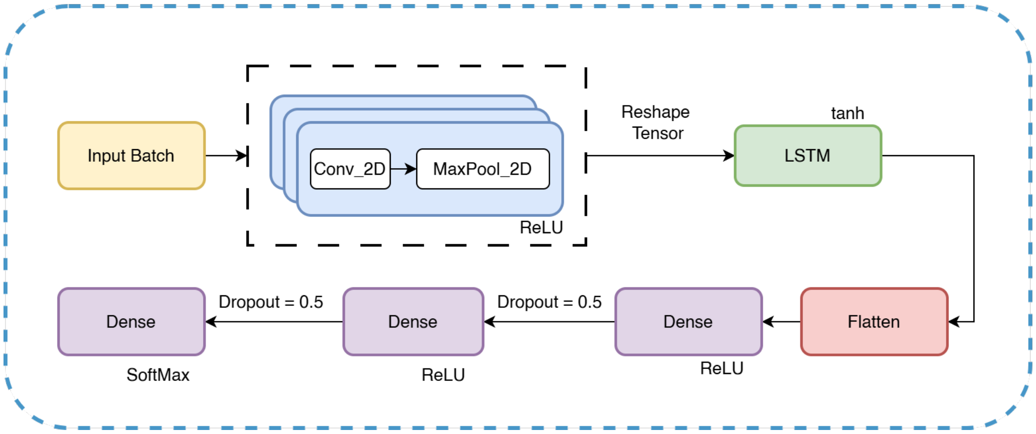

Fig. 2. Architecture of the CNN-LSTM Hybrid Model

a sequence of 2D convolutional and MaxPooling layers to extract unique traits from input images. Convolutional layers use filters to capture spatial hierarchies, with 32, 64, and 128 filters in each conv2D layer. MaxPooling layers reduce feature maps while keeping essential information, with a pool size of 2x2 and one MaxPooling layer after each conv2D layer. The activation function for conv2D layers is the Rectified Linear Unit (ReLU), providing non-linearity for understanding complex patterns.

Following the convolutional and pooling layers, the output is converted into an arrangement suitable for feeding into the LSTM layer. Typically, this procedure includes lessening the spatial dimensions while preserving the feature dimension. LSTM layers store information about temporal dependencies and excel in sequence prediction. These layers regulate output values with the 'tanh' activation function. Next, the flattening of the LSTM layer output generates an extended vector. LSTM output must be flattened to feed the Dense layers of 2D data. The final classification comprises dense layers based on CNN and LSTM parameters.

- First dense layer: 512 units, ReLU activation
- Dropout: Dropout minimizes overfitting in dense layers and prevents the model from evolving into too neuron-dependent, regularizing it. A 0.5 dropout rate denotes that 50% of neurons randomly get turned off during training.
- Second dense layer: 128 units, ReLU activation, and dropout are 0.5.
- Output dense layer: 4 units have been employed here. The SoftMax activation function produces a class probability distribution in the final dense layer. The classification task utilizes this layer's output.

Table I contains values of different parameters used during training the model.

V. Performance Analysis

To test each model's classification ability we opted for known and already used metrics such as confusion matrix, accuracy, precision, recall, and f1-score [25]. All these metrics helped to grab a better view of the models, without going into complexities of terms like true positive, false negative etc. We took the macro average, as this is a multi-class classification task, shown in table II.

TABLE I
Model Compilation Parameters

Parameter	Value
Optimizer	Adam
Metrics	Accuracy
Learning Rate	3e-4
Epoch	20
Batch Size	16
Loss Function	Categorical Cross-entropy

TABLE II
Performance comparison of All Models (Macro avg)

Architecture	Acc.	Pre.	Rec.	F1	Param (Million)
AlexNet	85	83	82	81	29.98
ResNet101 V2	89	91	88	90	43.74
MobileNet V2	90	93	90	91	2.9
CNN-LSTM	**94**	**94**	**94**	**94**	**55.4**

The experiment used AlexNet as the baseline deep CNN architecture, achieving an average accuracy of 85%. ResNet and MobileNet were used as comparison models. MobileNet V2 shows better accuracy with much fewer parameters. The Hybrid CNN-LSTM architecture achieved 94% accuracy, maintaining similar precision-recall and f1 scores. A higher parameter count helped extract the most from training, but there's still room for optimization. Lower parameter models are cost-effective and easy to deploy.

The predicted outputs were displayed in a normalized confusion matrix in figure 3. The Y-axis indicate test set classes. The true labels are COVID-19, NORMAL, PNEUMONIA, and TUBERCULOSIS. The X-axis holds the predicted classes in

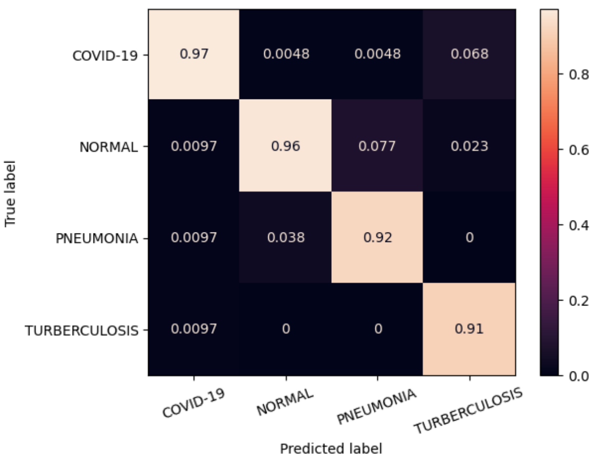

Fig. 3. Normalized Confusion Matrix of Test Set

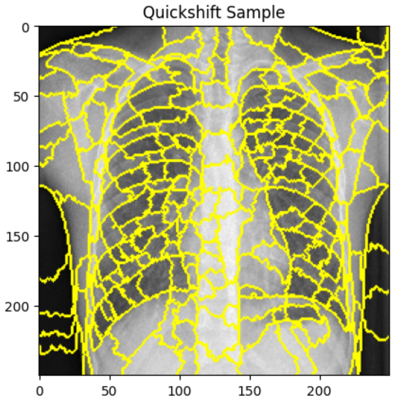

Fig. 4. Quickshift segmentation and superpixel formation

the matrix columns. The On-Diagonal Values are the correctly identified categories. For instance, the top-left cell (COVID-19 predicted as COVID-19) receives a value of 0.97, indicating 97% correct identification. The Off-Diagonal Values are misclassifications. The first row and the final column number 0.068 denotes that 6.8% of COVID-19 patients were miscategorized as TUBERCULOSIS. The confusion matrix depicts the class-wise performance of the model. TUBERCULOSIS had a 9% misclassification rate. Having class imbalance caused this issue. Also, the image quality significantly impacted the model's performance.

A. *LIME Analysis*

LIME works by applying quick-shift segmentation to form superpixels. An example is given in figure 4. When loaded into the LIME explainer, an image is divided into numerous segments based on the difference in brightness, colour, hues and saturation. LIME finds the segments important to the machine learning model, in this case, the CNN-LSTM Hybrid model.

Several samples of LIME have been represented in figure 5, 6 and 7. Each has three consoles. The first one shows the prediction, the second console shows the ”pros-cons” sections identified by LIME and lastly, a heatmap representation of LIME explanation. The LIME procedure begins by splitting the picture array into superpixels. These are continuous picture patches that have comparable hues and brightness. In the second console, the green regions are pros and the red are cons. On the third console, the red shades indicate these regions were important to the CNN-LSTM algorithm for this classification. The bluish regions indicate less weight to these areas.

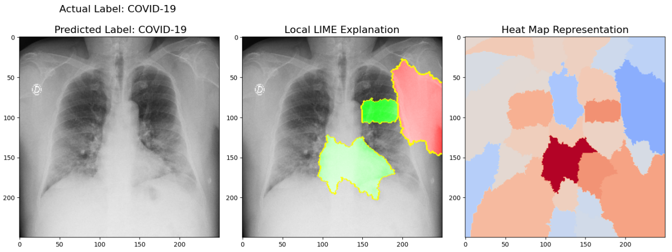

Fig. 5. LIME explanation on covid-19 sample

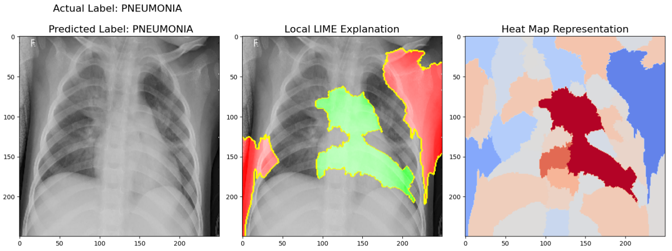

Fig. 6. LIME explanation on pneumonia sample

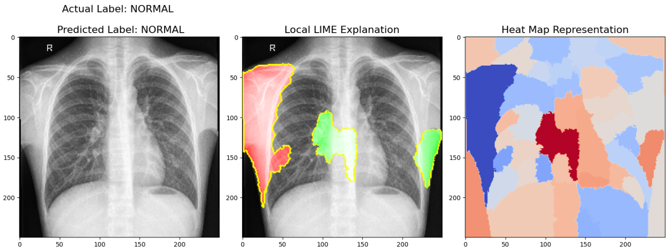

Fig. 7. LIME explanation on normal sample

In figure 5, the covid-19 case was identified correctly. The pros-cons area showed the green areas strongly correlated to this classification. The cons are generally the regions that are classified wrongly to another class. This is more vibrant from the heatmap, as it shows all the segmentation the LIME explainer model made while fitting it. Deep red means very important, and deep blue means least important to this prediction. Similar outputs are shown in figure 6 and 7.

So, LIME demonstrates which portions of the X-ray the model estimates, revealing its decision-making process. Consistent green highlights in the second console recommend that the model correctly recognise disease-related aspects.

VI. Conclusion

The LIME-Ray approach solves the transparency issue of Neural Networks while classifying multi-class diseases from X-ray images. We deployed the CNN-LSTM hybrid model for the classification, which achieved better performance compared to other CNN architecture and transformer models with pre-trained weights. This novel approach combines CNN with LSTM which tackles variability in image quality. This hybrid approach blends CNNs for spatial feature extraction with LSTMs for temporal pattern identification, boosting the model's ability and resilience. The predicted outputs are then explained using LIME. It shows the X-ray segmentation the model takes into account, revealing its decision-making process. The models' findings could assist researchers and practitioners apply XAI to assess COVID-19, Pneumonia, and Tuberculosis patients utilizing Chest X-ray images.

Acknowledgement

All the codes used in this experiment have been open-sourced on LIME-Ray Github Repository. The dataset is available at Chest X-Ray.

References

[1] V.-A. Surdu and R. Győrgy, "X-ray diffraction data analysis by machine learning methods—a review," *Applied Sciences*, vol. 13, no. 17, p. 9992, 2023.

[2] M. A. Mohammed, K. H. Abdulkareem, B. Garcia-Zapirain, S. A. Mostafa, M. S. Maashi, A. S. Al-Waisy, M. A. Subhi, A. A. Mutlag, and D.-N. Le, "A comprehensive investigation of machine learning feature extraction and classificationmethods for automated diagnosis of covid-19 based on x-ray images." *Computers, Materials & Continua*, vol. 66, no. 3, 2021.

[3] J. Greasley and P. Hosein, "Exploring supervised machine learning for multi-phase identification and quantification from powder x-ray diffraction spectra," *Journal of Materials Science*, vol. 58, no. 12, pp. 5334–5348, 2023.

[4] X. Zhao, Y. Luo, J. Liu, W. Liu, K. M. Rosso, X. Guo, T. Geng, A. Li, and X. Zhang, "Machine learning automated analysis of enormous synchrotron x-ray diffraction datasets," *The Journal of Physical Chemistry C*, vol. 127, no. 30, pp. 14 830–14 838, 2023.

[5] J. Rasheed, A. A. Hameed, C. Djeddi, A. Jamil, and F. Al-Turjman, "A machine learning-based framework for diagnosis of covid-19 from chest x-ray images," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, pp. 103–117, 2021.

[6] J. D. López-Cabrera, R. Orozco-Morales, J. A. Portal-Diaz, O. Lovelle-Enríquez, and M. Pérez-Díaz, "Current limitations to identify covid-19 using artificial intelligence with chest x-ray imaging," *Health and Technology*, vol. 11, no. 2, pp. 411–424, 2021.

[7] X. Ying, "An overview of overfitting and its solutions," in *Journal of physics: Conference series*, vol. 1168. IOP Publishing, 2019, p. 022022.

[8] C. F. G. D. Santos and J. P. Papa, "Avoiding overfitting: A survey on regularization methods for convolutional neural networks," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10s, pp. 1–25, 2022.

[9] A. Halder and B. Datta, "Covid-19 detection from lung ct-scan images using transfer learning approach," *Machine Learning: Science and Technology*, vol. 2, no. 4, p. 045013, 2021.

[10] A. Alqudah and A. M. Alqudah, "Sliding window based deep ensemble system for breast cancer classification," *Journal of Medical Engineering & Technology*, vol. 45, no. 4, pp. 313–323, 2021.

[11] S. Mehmood, T. M. Ghazal, M. A. Khan, M. Zubair, M. T. Naseem, T. Faiz, and M. Ahmad, "Malignancy detection in lung and colon histopathology images using transfer learning with class selective image processing," *IEEE Access*, vol. 10, pp. 25 657–25 668, 2022.

[12] C. Sitaula, T. B. Shahi, S. Aryal, and F. Marzbanrad, "Fusion of multi-scale bag of deep visual words features of chest x-ray images to detect covid-19 infection," *Scientific reports*, vol. 11, no. 1, p. 23914, 2021.

[13] M. K. Mahbub, M. Biswas, L. Gaur, F. Alenezi, and K. Santosh, "Deep features to detect pulmonary abnormalities in chest x-rays due to infectious diseasex: Covid-19, pneumonia, and tuberculosis," *Information Sciences*, vol. 592, pp. 389–401, 2022.

[14] N. N. Qaqos and O. S. Kareem, "Covid-19 diagnosis from chest x-ray images using deep learning approach," in *2020 international conference on advanced science and engineering (ICOASE)*. IEEE, 2020, pp. 110–116.

[15] C. Sitaula and M. B. Hossain, "Attention-based vgg-16 model for covid-19 chest x-ray image classification," *Applied Intelligence*, vol. 51, no. 5, pp. 2850–2863, 2021.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[17] M. M. Ahsan, K. D. Gupta, M. M. Islam, S. Sen, M. L. Rahman, and M. Shakhawat Hossain, "Covid-19 symptoms detection based on nasnetmobile with explainable ai using various imaging modalities," *Machine Learning and Knowledge Extraction*, vol. 2, no. 4, pp. 490–504, 2020.

[18] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[19] T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, T. Hamid, M. T. Islam, S. Kashem, Z. B. Mahbub *et al.*, "Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization," *Ieee Access*, vol. 8, pp. 191 586–191 601, 2020.

[20] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[22] R. Gao, R. Wang, L. Feng, Q. Li, and H. Wu, "Dual-branch, efficient, channel attention-based crop disease identification," *Computers and Electronics in Agriculture*, vol. 190, p. 106410, 2021.

[23] A. Souid, N. Sakli, and H. Sakli, "Classification and predictions of lung diseases from chest x-rays using mobilenet v2," *Applied Sciences*, vol. 11, no. 6, p. 2751, 2021.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[25] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," *arXiv preprint arXiv:2008.05756*, 2020.