Predicting Pneumonia from X-Ray Image

Miao Li, Yujie Jiang, Jariel Yang, Maxine Xu

May 24, 2023

Abstract

In this project, we aim to develop a deep learning algorithm to diagonalize pneumonia from chest x-ray images of one to five years old from Guangzhou Women and Children's Medical Center. After conducting an in-depth literature review, we performed image augmentation to obtain a balanced dataset and then used PyTorch to build a 3-layer convolutional neural network (CNN). To further improve the accuracy of our model, we performed a series of hyperparameter tuning, including different dropout rate, training epochs, learning rate, and additional layers. During the experiments, we customized a 3-layer model, which achieved an accuracy of 93.75% and a recall of 100%, and identified overfitting as our main challenge.

1 Model Construction

1.1 Data and Task Set Up

With the assistance of Python's os and pandas modules, we managed to calculate the counts of labels in the training, testing, and validation datasets provided by Kaggle. The results in Table 1 show that in the training data, we have 1,341 instances labeled as "Normal", and 3,875 instances labeled as "Pneumonia"; in the testing data, 234 instances are "Normal" and 390 instances are "Pneumonia"; both label counts of "Normal" and "Pneumonia" are 8 in the validation data.

Table 1. Label Counts

	Normal	Pneumonia	Total
Training Data	1341	3875	5216
Testing Datas	234	390	624
Validation Data	8	8	16

In addition to presenting the distribution of label counts, we further calculated the occurrences of different unique image sizes present in our data sets. As is revealed by the results in Table 2, we have 1,220 unique image sizes in the training data, among which 538 instances belong to the "Normal" and 682 instances belong to "Pneumonia"; the testing data comprises 328 unique image sizes, including 195 "Normal" and 133 "Pneumonia"; there are 6 "Normal" and 8 "Pneumonia" in the validation data.

Table 2. Counts of Image Sizes

	Normal	Pneumonia	Total
Training Data	538	682	1220
Testing Datas	195	133	328
Validation Data	6	8	14

Class imbalance is evident within the training data as 74.3% of the images are classified as pneumonia. We found that many papers implemented transformations not only to facilitate comparisons but also to address the challenge of imbalanced data by augmenting the datasets/images (e.g., Trivedi and Gupta, 2022; Bharati et al., 2021). However, we noticed that few papers provided explicit details on the selection process; they merely presented their final choices. Nevertheless, one paper by Giełczyk et al. (2022) conducted a comparison of various transformations and concluded that combining ”histogram equalization” and ”gaussian blur” yielded high accuracy. Therefore, when tripling the normal images within the dataset, we incorporated this combination along with other transformations into our data augmentation step to gain a training dataset with 50.94% normal images and 49.06% pneumonia images.

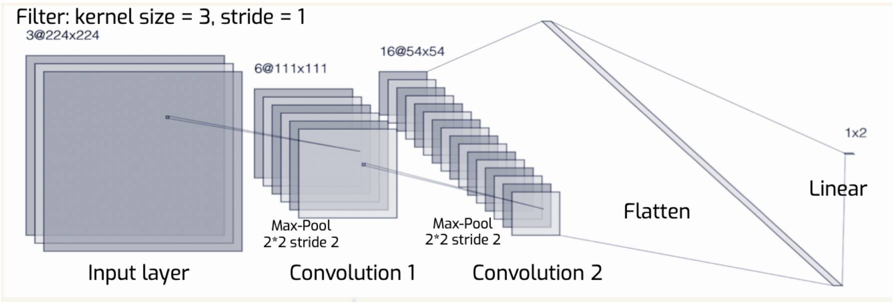
At this stage, we opted to build our “Dataset” class based on pandas data frames, where each “Dataset” class is initialized by loading it into a corresponding pandas data frame, such as a training data frame. The construction strategy that we made here aligns with our preliminary data pre-processing choices where we created pandas data frames to store and manipulate the data. We believe this will not only facilitate a smoother and more consistent data loading process, but also provide a solid foundation for future data management and analysis by ensuring our pipeline remains scalable and adaptable to potential modifications or expansions as the project progresses. This, in turn, contributes to a more robust and reliable deep learning process in the long run.

1.2 Model structure

Convolutional Neural Networks (CNN) is a dynamic learning algorithm primed to identify visual features within images. By adjusting weights and biases as needed, it effectively discerns different objects from various aspects of an image. This capability to extract hidden information makes CNN exceptionally efficient in visual data analysis and object recognition tasks, surpassing traditional classification algorithms. Additionally, CNN is renowned for its low pre-processing requirements but extensive filtering process, which enables more comprehensive learning of characteristics.(Khan et al., 2021). Since CNNs have been proven effective in handling high-dimensional data, capturing spatial relationships in images and automatically extracting hierarchical features, our deep learning model leverages the advantages of CNNs to predict pneumonia using X-ray images.

Inspired by the classic LeNet model, our work began with constructing a three-dimensional network architecture from scratch. This novel network comprises two convolutional layers and one linear layer. Each convolutional layer is enhanced with a batch normalization layer, followed by an activation layer, and a pooling layer. This architecture was motivated by the need for a model capable of extracting low-level features such as edges and textures (first convolutional layer), and high-level features like the specific structure of pneumonia-infected lung regions (second convolutional layer).

To prevent the model from overfitting to the training data, an extra dropout layer was integrated into the network. This layer introduces randomness into the model during training, forcing the neurons to be more robust and hence improving generalization to unseen data. The inclusion of the dropout layer mitigates the risk of over-reliance on any single feature or neuron, encouraging a more distributed and hence more reliable learning process.

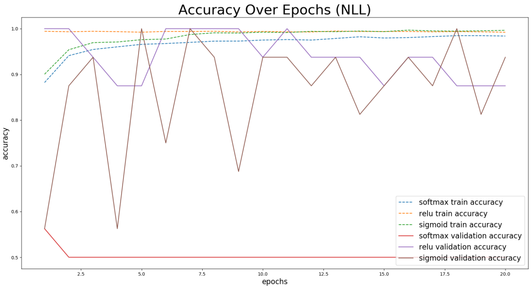


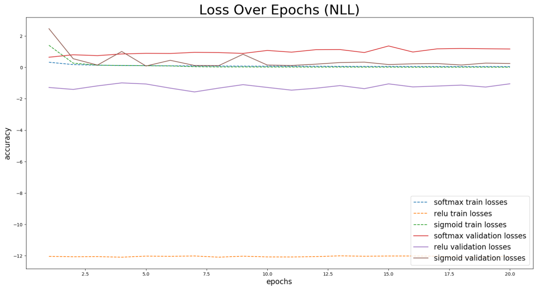
Since we are creating an architecture from scratch, we have the flexibility to experiment with various aspects of our model. This includes varying the loss functions and activation layers to optimize the performance. By adjusting these functions and layers, we can manipulate the way our model learns and responds to the data. Further, we can increment or decrement the network dimensions to study the impact on model complexity and performance.

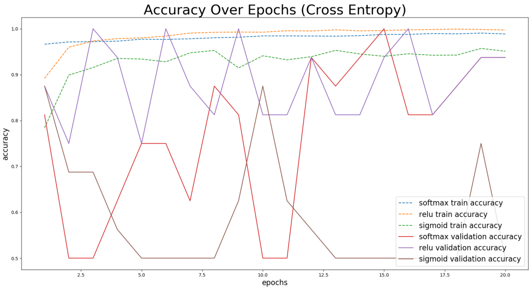
In the following sections, we will delve into the training and testing processes of our model. Through our analysis, we aim to provide insightful findings, and discuss potential improvements and applications. We hope our work contributes to the evolving landscape of machine learning techniques applied in healthcare, paving the way towards more accurate and efficient disease prediction models.

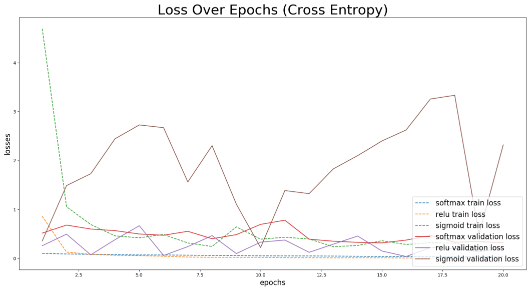
1.3 Hyperparameter Tuning

Throughout this project, we underwent multiple iterations of our neural network model, fine-tuning its hyperparameters to optimize prediction accuracy on the validation data. First, we constructed a neural network with two convolutional layers and one fully connected layer. In order to decide our loss and activation functions, we ran two groups of comparison with 20 epochs each. Both groups compared the performance of the ReLu, sigmoid, softmax activation functions with the first group utilizing the cross entropy loss function and negative log likelihood in the second.





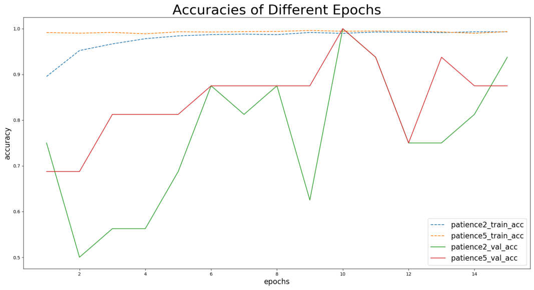


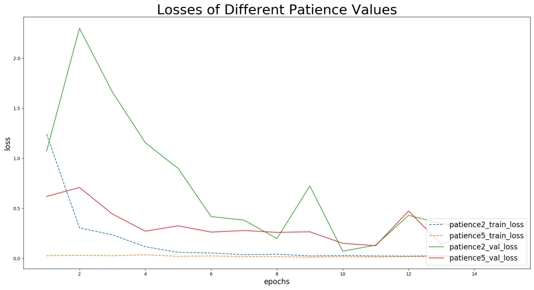


From the above graphs it can be seen that the accuracy increased with the negative log likelihood as the loss function. The model with sigmoid as activation function especially showed great improvement in the accuracy, yet the model with ReLU still has a more stable and accurate performance across epochs. This finding supports Qing Li's finding that ReLu improves both learning speed and classification performance in CNN applications (Li et al. 2014), especially when with larger input size. Compared to sigmoid, Li concludes that the ReLU activation function improves classification performance by 2.5% and the network converges much faster. At the same time, we have observed that our models tend to overfit into the training data, featuring very high training accuracy and relatively low validation accuracy. We therefore decided that our principle for the picking a better model will be choosing a simpler model when accuracy and loss are similar.

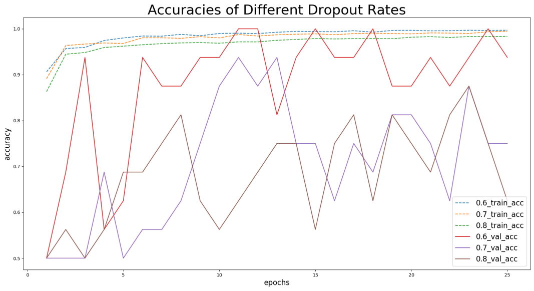
Hence, for the second of hyperparameter tuning, we start with the following baseline model: the first two convolutional layers in our preliminary model have 6 and 16 filters, with a kernel size of 3x3, a stride of 1, and no padding. Each convolutional layer is followed by a batch normalization layer, ReLu activation, dropout regularization, and max pooling with a kernel size of 2x2 and a stride of 2.

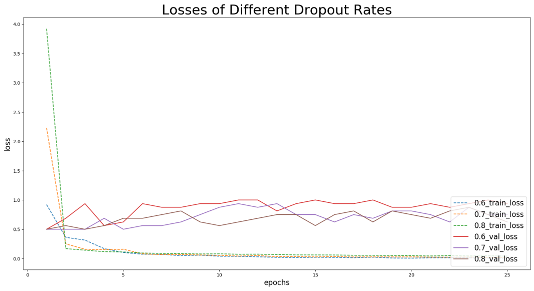
In the Accuracy Over Epochs (NLL) graph, the training accuracy ReLu illustrates that the model is almost perfectly predicting cases of pneumonia in the training dataset, which contrasts the fluctuating the validation accuracy. In order to prevent overfitting, we decreased our starting learning rate from 0.01 to 0.001 and utilized a learning rate scheduler. The graphs below showcase the results from testing a patience level of 2 and 5. Looking at the validation accuracy and loss, the patience level of 5 has a higher accuracy across most epochs, illustrating that a higher patience level provides more stability during training and helps our model tolerate fluctuations in the validation performance before making premature adjustments to the learning rate.


Accuracies of Different Epochs


Losses of Different Patience Values

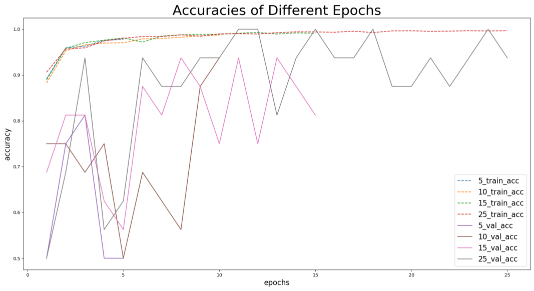
The inclusion of dropout is another measure we implemented to prevent overfitting. In "Heart Disease Prediction Using Modified Version of LeNet-5 Model", Shaimaa Mahmoud discovered that the inclusion of max pooling and dropout regularization enhanced their LeNet model's accuracy in detecting heart disease from 89.24% to 98.38%, which inspired us to include these two operations (Mahmoud 2022). Since dropout acts as a form of regularization by adding noise to the network, we tested 0.6, 0.7, and 0.8 dropout levels and graphed the accuracies in the graph below.

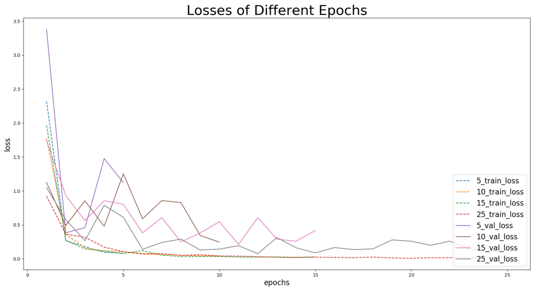

Accuracies of Different Dropout Rates


Losses of Different Dropout Rates

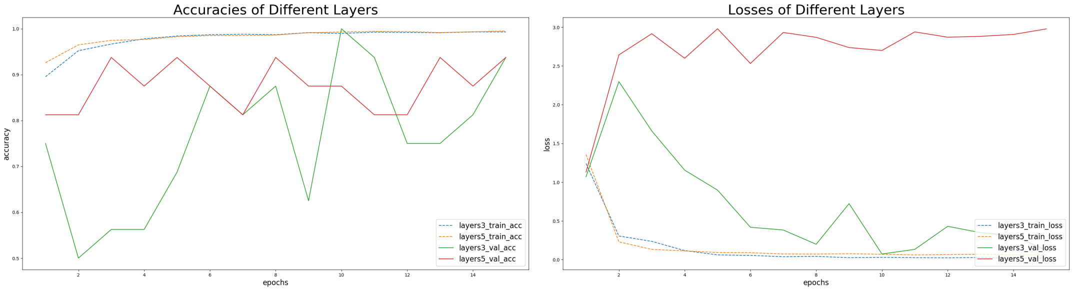
Because the 0.6 dropout level achieves the highest validation accuracy, we adopted it within our model.

In terms of epochs, we chose to train our model using 15 epochs as it was both time efficient and achieved a high level of validation accuracy.


Accuracies of Different Epochs


Losses of Different Epochs

Lastly, we tested our three layer neural network with a five layer network, which had two convolution layers and three linear layers with the same hyperparameters.



We expected the five-layer model to improve the model by capturing more detailed features and indeed we found that in general the five layer model has stable accuracy and loss trends. Although the validation accuracy is shown to be higher in general for the five-layer model, we ultimately decided to choose the three-layer model as the training accuracy for the fiver-layer model was higher than the three-layer model, which indicates a higher chance of overfitting. Additionally, the loss for the three layer model has a negative trend, which is an indication that the model is gradually improving and converging towards the best solution.

1.4 Assessing Classifier Accuracy

A consistent trend observed in studies focused on pneumonia detection, such as the work of Rohit Kundu et al. titled ”Pneumonia detection in chest X-ray images using an ensemble of deep learning models,” is the evaluation of classifier accuracy using four key metrics: accuracy, precision, recall, and the F1-score (Kundu, 2021). These metrics are widely employed in various studies, and we have also incorporated them into our project. Specifically, we have utilized the accuracy measure to guide our selection of activation and loss functions, while the remaining three metrics have been employed to assess our model's performance on the test dataset.

Overall, our model achieved an accuracy of 93.75%, a recall rate of 100%, a precision of 89%, and an F1 score of 94%. Notably, the 100% recall rate signifies that our model successfully identified all positive instances within the test set, leaving no positive cases undetected. This capability is of utmost importance in medical diagnostics, particularly in our case of pneumonia detection, as it ensures that no patients are overlooked.

Nevertheless, it is crucial to consider other evaluation metrics such as precision, accuracy, and F1 score, as a high recall rate may lead to increased false positives or misclassifications in general. In our model, the accuracy rate of 93.75% demonstrates that it correctly predicts approximately 93.75% of all instances in the dataset, indicating a minimal number of misclassifications. Furthermore, the precision of 89% suggests that roughly 89% of instances classified as positive by the model are indeed positive cases. Lastly, the F1 score of 94% indicates a favorable balance between precision and recall, taking into account the model's ability to identify positive instances and the accuracy of its positive predictions.

2 Model Analysis

2.1 Working with black-and-white

Black-and-white images, including X-ray imagery, bring a unique set of benefits and challenges to the table. On the plus side, these grayscale images are less complex than their color counterparts due to their singular channel, in contrast to the three channels (RGB) usually found in color images. This simplicity reduces computational demands, making grayscale images a more accessible choice for many tasks. Additionally, grayscale images often exhibit heightened contrast between different image areas. This is particularly advantageous in medical imaging, where grayscale can enhance the distinguishability of certain features, potentially aiding diagnosis.

However, working with grayscale images is not without its challenges. A key issue is the loss of color information, which can be instrumental in recognizing certain patterns or features within an image. Additionally, the use of shades of gray can sometimes make it difficult to differentiate between features with similar grayscale values, creating ambiguity. This issue is exacerbated when analyzing X-ray images, which frequently contain noise and artifacts that can obstruct analysis. These can range from physical scratches or dust to more complex artifacts intrinsic to the X-ray imaging process. Another challenge arises from normal anatomical variations between individuals. It can be difficult to build a model that accurately identifies diseases such as pneumonia across a diverse range of individuals, thereby presenting a hurdle in the effective application of grayscale imaging.

2.2 Potential application and limitations of the model

According to a study done by Henry Ford Health System, 72 percent of patients are misdiagnosed with pneumonia upon readmission (Henry Ford Health System, 2010). Although our model achieved an accuracy rate of 93.75 percent in correctly identifying patients with pneumonia in the test dataset, there are limitations to consider when assessing its applicability in a clinical setting. The external validity of our model is uncertain, as our data is restricted to X-ray images of one to five-year-old children from a hospital in Guangzhou, China (Kermany, 2018). This restricts the generalizability of our model to vulnerable populations such as African Americans, smokers, patients with lung disease, and those admitted through the Emergency Department, as highlighted by the Henry Ford Health System study (Henry Ford Health System, 2010).

Furthermore, despite efforts to prevent overfitting through the introduction of dropout and simplifying the model architecture to three layers, our model still struggled with overfitting to the training and validation data. Since our training, testing, and validation data originate from the Kaggle dataset, it is likely that our model would perform less accurately when applied to different populations.

The findings from the Henry Ford Health System study also highlight the presence of bias in healthcare data, with African American patients being twice as likely to be misdiagnosed compared to Caucasian patients. If training data is unrepresentative or biased in terms of demographics and socio-economic variables, the model's predictions are likely to be skewed and discriminatory, reflecting this "trash in, trash out" sentiment.

In the field of pneumonia detection using X-ray images, notable models such as LeNet, ResNet-18, and DenseNet-121 have been widely used (Kundu, 2021). In the study "Pneumonia detection in chest X-ray images using an ensemble of deep learning models," Kundu aggregates the predictions from GoogLeNet, ResNet-18, and DenseNet-12 to create an ensemble model. This ensemble model was effective in reducing bias by averaging predictions and smoothing out an individual model's errors, increasing model diversity, and decreasing overfitting. In relation to our model, the high number of layers of these three models and the computation cost to train all three of them before combining them presents a second limitation of our model. X-ray images contain spatial patterns that may be too complex for our model to capture. Because of the computational complexity and time-consuming nature of training such high layer models, our 3 layer model has limited flexibility in capturing the spatial context of the x-rays as well as a limited capability for feature extraction.

In the future, we hope to explore the following algorithms that are emerging in the realm of computer vision. These algorithms include but are not limited to Transfer Learning, Object Detection, Semantic Segmentation, Instance Segmentation, Anomaly Detection, etc.(Tian et al., 2022, Pang et al., 2021). Transfer Learning, for instance, is an algorithm that is widely used in neural networks. It makes use of a concept akin to human cognition, where previously acquired knowledge is applied to new tasks. Networks are trained and tested on diverse datasets, and this acquired knowledge is then leveraged to train and test new datasets—a process emulating cognitive transfer of learning. Keras offers a range of transfer learning models such as Xception, VGG16, VGG19, ResNet, Inception-v3, MobileNet, DenseNet, and NASNet, all previously trained on expansive datasets. These models can be fine-tuned to specific tasks with smaller data volumes, embodying the principle of transfer learning(Jain et al., 2020). In addition to Transfer Learning, Object Detection and Semantic Segmentation also play a crucial role in medical image processing, where the former focuses on identifying lesion locations and classifying various objects in medical imaging; the other is used to segment images based on semantics and pixel information(Yang et al., 2021).

Lastly, our methodology for fine-tuning parameters involved adapting hyperparameters that resulted in the highest validation accuracy. The order in which we adjusted parameters followed a specific sequence, starting with the activation function, loss function, learning rate, dropout rate, epochs, patience, and layers. Some hyperparameters, like learning rate and patience, are interdependent, influencing the optimal values of other hyperparameters. In future work, exploring different orders of parameter selection could potentially lead to the development of a more accurate pneumonia classification model.

In conclusion, while our model demonstrated promising accuracy on our specific dataset, limitations related to external validity, bias, computational complexity, and limited flexibility in capturing spatial context and feature extraction should be acknowledged. Further research and improvements are necessary to develop a more robust and generalizable model for pneumonia detection in chest X-ray images.

3 Discussion of Implication

Applying predictive algorithms in diagnostic imaging means both opportunities and challenges. As the distribution of algorithms takes very little time and monetary cost, an accurate diagnostic algorithm means one big step to universal access to essential medical resources, which will especially benefit the more disadvantaged population facing burdens (like money or transportation resources) to gain traditional medical service. Having algorithms in image-based diagnosis is thus crucial for expanding the application of automatic tools in the medical sphere as it is more complicated than just comparing numerical biomarkers to a certain threshold.

Moreover, in the case of diagnosing pneumonia in children, there is a current lack of consensus over the definition and classification of pneumonia (Mackenzie, 2016) and many clinical diagnoses are currently based on research of adult samples (Rodrigues Groves, 2018). So from a research perspective, introduction of deep learning would potentially help identify x-ray image features that are specific to the children and thus inspire medical research.

However, there are also challenges associated with employing computer vision algorithms on a large scale, including problems associated with sampling and bias as well as the need for relative policy and laws. To begin with, the accuracy of the algorithm depends on the sampling of training data and many factors including age and vaccination status may play a part in diagnosis. So just as stated in the limitation part, underrepresented groups in the sample would likely suffer from more misdiagnosis due to a lack of data for the algorithm to pick useful features specific to the groups. By the same token, any bias in doctors' diagnosis would also introduce systematic bias in the algorithm. Due to the lack of consensus over the definition and that pneumonia shares similar symptoms with many other respiratory diseases (Stokes et al. 2021), doctors adopting different standards would cause clinical biases that are unknown to algorithm developers.

Meanwhile, using machine learning algorithms to detect disease still faces medicolegal challenges and there are regulatory gaps that would need to be bridged (United States Government Accountability Office, 2022). For example, the standard and evidence needed for FDA's review and approval as well as accountability for misdiagnosis caused by algorithms. Therefore, it requires both effort from the medical sector and the policy sector before machine learning algorithms can be adopted in real-life settings.

References

Giełczyk, Marciniak,Tarczewska and Lutowski (2022). Pre-processing methods in chest X-ray image classification. Plos one, 17(4), p.e0265949.

Health Ford Health System (2010). “Pneumonia Often Misdiagnosed on Patient Readmissions, Studies Find.” ScienceDaily, 28 Oct. 2010, www.sciencedaily.com/releases/2010/10/101022123749.htm.

Jain, R., Nagrath, P., Kataria, G., Kaushik, V. S., Hemanth, D. J. (2020). Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning. Measurement, 165, 108046.

Kermany et al (2018). “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning.” Cell, 22 Feb. 2018, pubmed.ncbi.nlm.nih.gov/29474911/.

Khan, A. A., Laghari, A. A., Awan, S. A. (2021). Machine learning in computer vision: a review. EAI Endorsed Transactions on Scalable Information Systems, 8(32), e4-e4.

Kundu et al. (2023) “Pneumonia Detection in Chest X-Ray Images Using an Ensemble of Deep Learning Models.” PLOS ONE, doi.org/10.1371/journal.pone.0256630.

Li et al. (2014) ”Medical image classification with convolutional neural network,” 2014, 13th International Conference on Control Automation Robotics Vision (ICARCV), Singapore, pp. 844-848, doi: 10.1109/ICARCV.2014.7064414.

Mackenzie (2016) The definition and classification of pneumonia. pneumonia 8, 14 .
https://doi.org/10.1186/s41479-016-0012-z

Mahmoud (2022). “Heart Disease Prediction Using Modified Version of Lenet-5 Model.” International Journal of Intelligent Systems and Applications, vol. 14, no. 6, 2022, pp. 1–12.,
https://doi.org/10.5815/ijisa.2022.06.01.

Pang, G., Shen, C., Cao, L., Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. ACM computing surveys (CSUR), 54(2), 1-38.

Rodrigues and Groves (2018), Community-Acquired Pneumonia in Children: the Challenges of Microbiological Diagnosis, Journal of Clinical Microbiology, doi: https://doi.org/10.1128/JCM.01318-17

Stokes et al. (2021), “A machine learning model for supporting symptom-based referral and diagnosis of bronchitis and pneumonia in limited resource settings”, Biocybernetics and Biomedical Engineering, doi: https://doi.org/10.1016/j.bbe.2021.09.002

Tian, D., Han, Y., Wang, B., Guan, T., Gu, H., Wei, W. (2022). Review of object instance segmentation based on deep learning. Journal of Electronic Imaging, 31(4), 041205-041205.

Trivedi and Gupta (2022), “A lightweight deep learning architecture for the automatic detection of pneumonia using chest X-ray images”. Multimedia Tools and Applications, 81(4), pp.5515-5536.

United States Government Accountability Office (2022), “Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics”,
https://www.gao.gov/assets/gao-22-104629.pdf

Yang, R., Yu, Y. (2021). Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. Frontiers in oncology, 11, 638182.