Application of Convolutional Neural Networks (CNNs) Reduces Radiological Diagnostic Errors in Classifying Pulmonary Diseases Based on Chest X-rays

Radha Munver[1*], Nidhi Parthasarathy[2], Sanya Badhe[2]

[1] Dwight-Englewood, Englewood, 07631, New Jersey, USA.
[2] Lynbrook High School, San Jose, 95129, California, USA.

[*]Corresponding author. E-mail(s): rmunver25@d-e.org;
Contributing authors: nidhi.parthasarathy@gmail.com; sanya.badhe@gmail.com

Abstract

Approximately one billion radiologic examinations are performed each year, with 3-5% (40 million) of diagnoses by radiologists being prone to error [1]. In the current era of clinical medicine, natural human error often results in misdiagnosis upon the review of X-rays, CT (computed tomography), and MRI (magnetic resonance imaging) scans. As a result, numerous patients receive incorrect diagnoses and may undergo treatment plans unsuited to their medical condition. Additionally, radiologists and treating physicians may encounter unfortunate adverse repercussions associated with allegations of negligence or malpractice. This study aims to address this shortcoming by examining image classification for chest diagnoses. In this study, the applications of Res-Net18, a pre-trained Convolutional Neural Network (CNN), to classify chest X-rays into 14 distinct chest diagnoses were explored. Leveraging the ChestMNIST dataset [2], computer vision techniques were used to discern the chest diagnoses based on the pre-provided labels using multi-label binary-class classification. The data was split as follows: 70% for training, 10% for validation, and 20% for testing. Sigmoid and binary cross-entropy with logits was used as the loss function for its accompanying optimizer and its outperformance of softmax with cross-entropy loss. The Res-Net18 model was high-performing, with an accuracy of 94.7% and Area Under Curve (AUC) of 0.772. Accuracies across the classes ranged from 86% to 99.8% (93.92% average). There were high precision scores for all of the classes with a micro average of 0.947, a low recall of 0.56, and an F1-Score of 0.54. This study has significant implications for the advancement of automating radiological diagnoses as an ambient intelligence technology and secondary confirmation for reducing errors and discrepancies, thereby facilitating more accurate diagnoses and enabling appropriate treatment.

Keywords: Convolutional Neural Network (CNN), ChestMNIST, Res-Net18, ambient intelligence, deep learning, sigmoid and binary cross-entropy with logits

1 Introduction

In 2022, the Kaiser Family Foundation reported that only 48.07% of primary care needs were met [3], underscoring a significant deficiency in healthcare accessibility. The growing physician shortage, compounded by increasing patient demand for medical care, has further exacerbated this issue. As a result, patients frequently face extended wait times for medical attention, a matter of heightened importance, particularly when achieving a prompt diagnosis is imperative.

This study addresses the aforementioned issues through the use of diagnostic imaging which plays a crucial role in the early detection and diagnosis of diseases. Radiological imaging has revolutionized the approach that hospitals take toward diagnosing and treating patients. Various imaging modalities are available and routinely performed at most hospitals. This accessibility suggests that a model capable of accurately detecting diseases through radiologic imaging would greatly enhance the process of identifying and treating diseases before they progress.

Our model can accurately identify and pinpoint 14 different diagnoses from chest X-ray images, thus reducing the wait time for radiologist interpretation. Furthermore, this model could serve as an ambient intelligence technology, providing a second opinion and validating radiologists' diagnoses, thereby reducing inadvertent errors, improving confidence, and enhancing the efficiency of treatment plans.

While this model may be a valuable tool in the healthcare industry, it is important to note that it is not intended to replace medical professionals, rather its intent is for it to be used as a tool to complement their expertise and assist them in making informed and timely decisions.

2 Methods

2.1 Dataset

The ChestMNIST dataset is derived from MedMNIST, a collection of medical images in the format of the Modified National Institute of Standards and Technology database. ChestMNIST is based on the NIH-ChestXray14 dataset, which comprises 112,120 labeled frontal-view X-ray images across 30,805 unique patients [4]. Of the 112,120 images in the dataset, 78,468 (70%) were allocated for training, 11,219 (10%) for validation, and 22,433 (20%) for testing. Data was converted from grayscale images (originally with dimensions of 28 x 28 pixels) into RGB (Red, Green, Blue) images (3 x 28 x 28 pixels) as shown in Figure 1. By adding the RGB dimension to the images, they could then be inputted into the Res-Net18 pre-trained model.

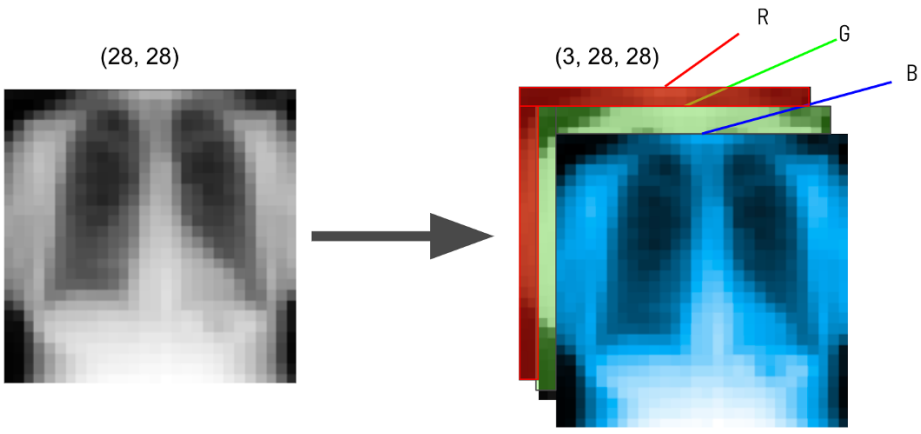

Figure 1: Resizing dimension of original grayscale X-rays from (28 x 28) to (3 x 28 x 28) by adding RGB dimension.

2.2 Modeling

Figure 2 delineates the procedural steps for the model. The first step involved splitting the dataset into the training and test sets, which then required feature selection and training. Training required several hours, after which trained classifiers were used, followed by evaluation and analysis. The approach to X-ray classification is demonstrated in Figure 3. The training and test sets were crucial in allowing the quantification of test errors and computation of metrics including accuracy, precision, and F1-score.

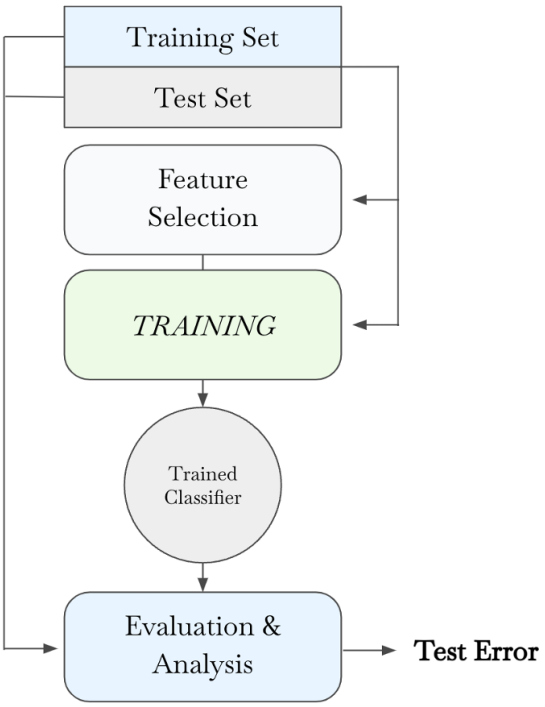

Figure 2: Workflow implementing the Res-Net18 framework.

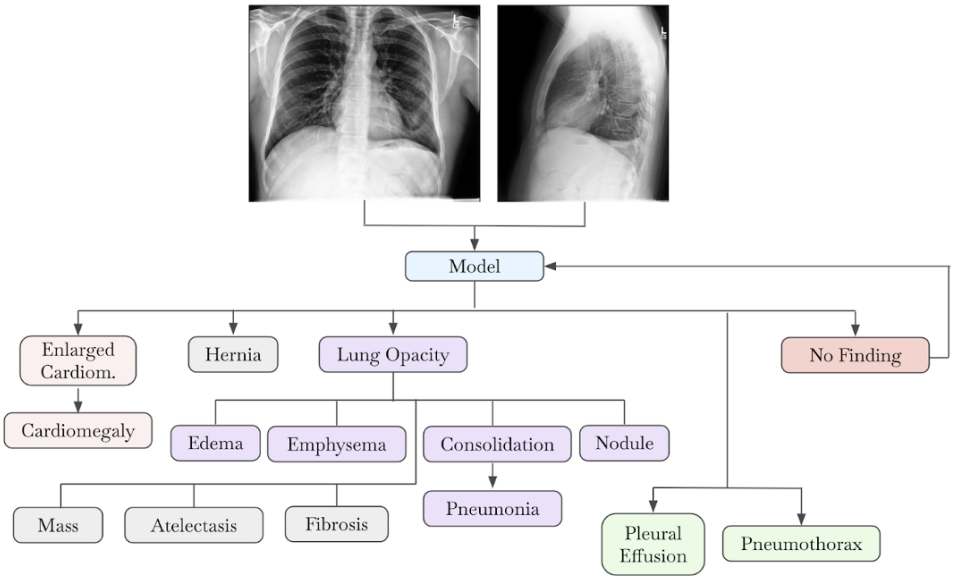

Figure 3: Process of X-ray classification from an image inputted into the model for assignment to one of the 14 diagnoses. An image that cannot be classified (i.e., "No Finding"), is re-inputted into the model.

2.3 Neural Network: Res-Net18

The Res-Net18 model is a CNN, a type of deep learning algorithm suited to analyze 3-dimensional visual data, containing a depth of 18 layers. The X-rays were originally formatted as 2-dimensional grayscale images. As the Res-Net18 model requires an RGB image as the input, it was necessary to convert the images from grayscale to RGB.

2.4 Training, Runtime, Validation, and Loss

The model was trained for 100 epochs using 2 GPUs (Graphics Processing Units). Our training model used two passes: a forward pass and a backward pass. The forward pass informed the model of any inaccuracies, while the backward pass notified the model of how to learn from the training data, enabling it to correct itself. As the model runtime was approximately 7 hours, TQDM progress bars were leveraged to indicate the progress of the model during the training process. Finally, training and validation losses were monitored over the epochs.

3 Results

3.1 Training & Validation Loss

The Python visualization library, Matplotlib, was used to track how the training loss and validation loss changed over time. As illustrated in Figure 4, the training loss exhibited a steady negative progression, decreasing at a relatively constant rate, indicating positive model performance. However, the validation loss was clearly more stochastic and varied, thus demonstrating overfitting of the model. This finding was a product of a small number of positive cases and a large number of negative cases, which is common in the medical field.

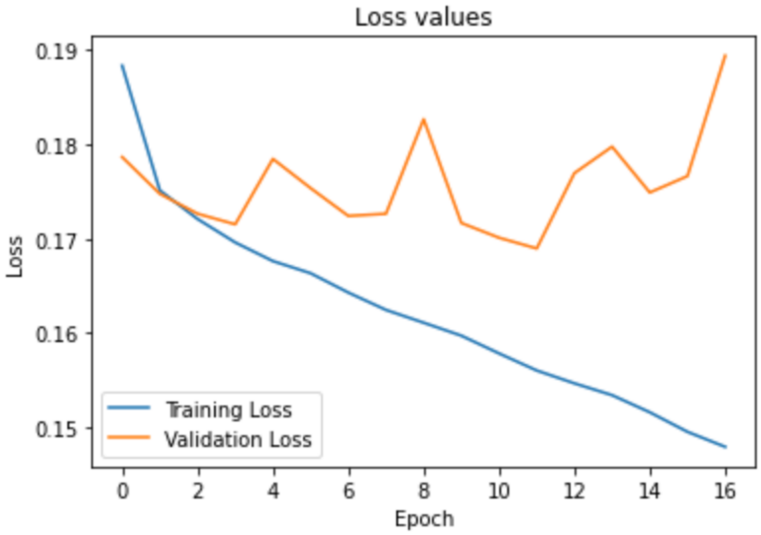

Figure 4: Training loss (blue) and validation loss (orange) graphed across epochs 0–16 during training.

3.2 Errors

Several challenges were encountered during the development of this study. For instance, the model demonstrated overfitting during training, which resulted in the validation loss increasing while the training loss decreased. Additionally, extensive time was spent in flattening the dataset through attempting to add it to a linear model. However, it was determined that this time consumption did not add a significant benefit to the project. Given that multi-label, binary class classification is necessary for an error-sensitive field such as medicine, it was essential to leverage loss functions with accompanying optimizers. Initially, softmax with cross-entropy loss were used; however, these were subsequently replaced with sigmoid and binary cross entropy with logits.

3.3 Accuracy

Accuracy across all the classes ranged from 86% to 99.8% (Figure 5). When these accuracies are plotted on a column chart (Figure 6), it is evident how closely the accuracies were aligned with each other, with a 93.92% average across all of the classes. This result indicates that the training data closely resembled the actual results, reflecting the commendable accuracy of the model. The next step was to quantify additional metrics to assess the model's overall performance.

Disease	Accuracy	Disease	Accuracy
1. Atelectasis	87%	8. Pneumothorax	94%
2. Cardiomegaly	97%	9. Consolidation	95%
3. Effusion	86%	10. Edema	98%
4. Infiltration	79%	11. Emphysema	97%
5. Mass	94%	12. Fibrosis	98%
6. Nodule	94%	13. Pleural	97%
7. Pneumonia	99%	14. Hernia	99.8%

Figure 5: Comparison of computed accuracies for each class label (diagnosis).

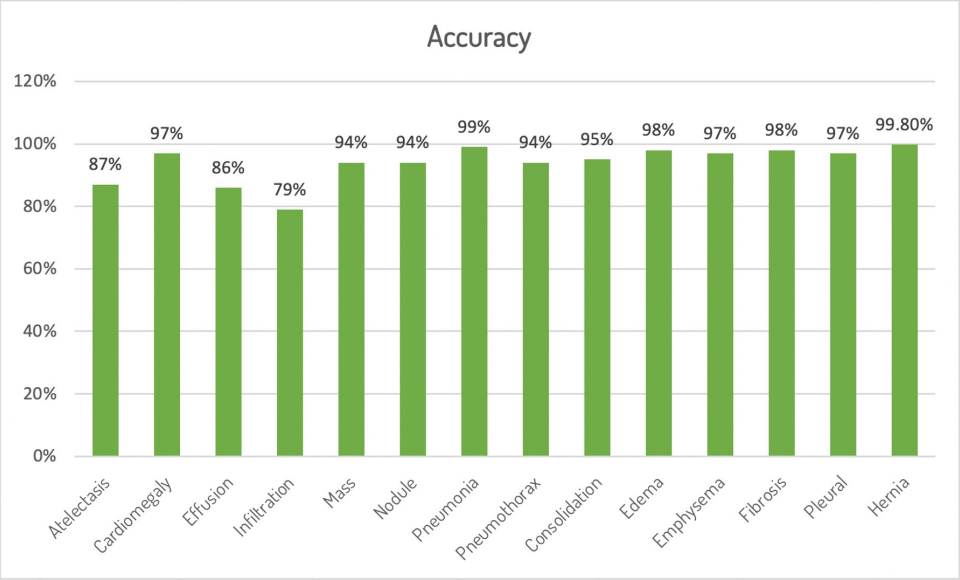

Figure 6: Computed accuracies for each class label (diagnosis). Average accuracy (purple) of 93.92% is displayed by dashed-line.

3.4 Precision

Figure 7 displays all of the class-wise precisions graphed on a column chart for analysis. It is notable that there is a relatively high precision for all of the classes with a micro average of 0.947. These results are encouraging, as improved performance with values closer to 1 correlate with superior precision.

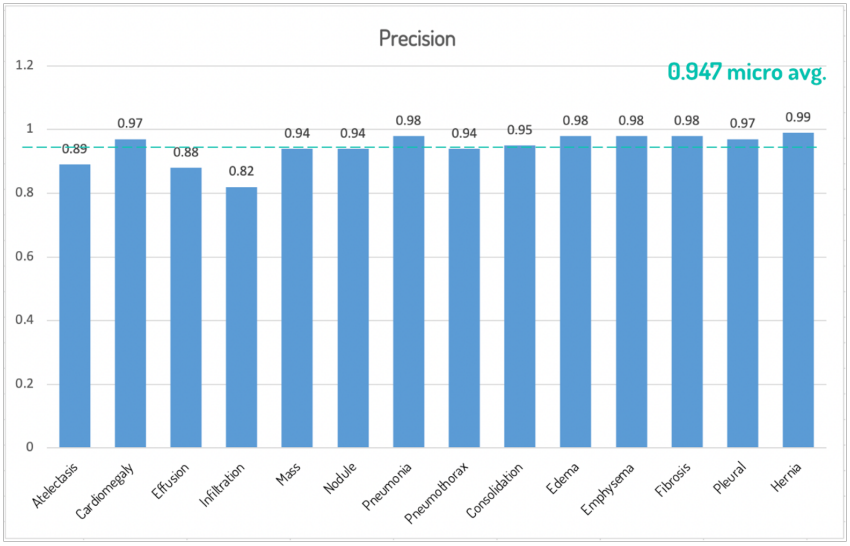

Figure 7: Computed precision for each class label (diagnosis). Micro-average precision (teal) of 0.947 displayed by dashed-line.

4 Discussion

4.1 Micro-Average Metrics

Our model demonstrated impressive precision, achieving a score of 0.95, highlighting its ability to accurately predict positive labels. Nonetheless, the associated recall of 0.56 and F1-score of 0.54 indicate that the model had difficulty in capturing all positive instances effectively. The diminished recall and F1-score suggest a potential inclination toward false positives, underscoring the need for additional refinement to attain a more balanced and thorough classification performance.

4.2 Area Under Curve (AUC)

The AUC, ranging from 0 to 1, is an important metric for assessing the effectiveness of our model, specifically due to its balanced nature. The AUC refers to the area beneath the ROC (Receiver Operating Characteristic) curve. The ROC graph illustrates the performance of our classification model across all classification thresholds. The two parameters that are used are the true positive rate and the false positive rate. Our AUC of 0.772, which is on the higher end of the spectrum, indicates that our model achieved a favorable result.

4.3 Confusion Matrices

The confusion matrix is comprised of four main aspects: true positive, true negative, false positive, and false negative. In this context, a positive sample indicates the presence of an abnormality, while a negative sample is associated with a normal chest X-ray.

The true positives and true negatives are almost evenly distributed among the confusion matrices, resulting in high precision and low recall. There are fewer positives compared to negatives (i.e., in

cardiomegaly, there are approximately 100 positives compared to 26,000 negatives) as illustrated in Figure 8. The difference between negatives and positives, in addition to the high standard deviation of the positives in the dataset, resulted in the inability to achieve a high recall. There were a large number of negatives in the dataset, creating potential bias in the model due to the imbalance of the positives and negatives. Due to the small number of positive cases, despite the high accuracy, the recall remained low, thus lowering the F1-score. This emphasizes the importance of considering all of the metrics, not just accuracy.

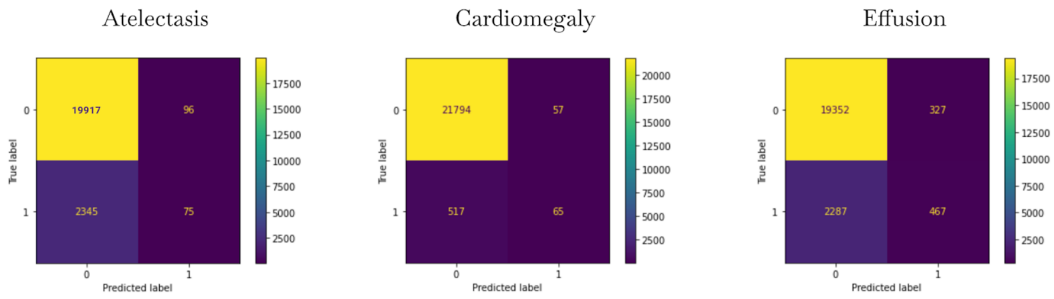

Figure 8: Sample confusion matrices for 3 diagnoses: atelectasis, cardiomegaly, and effusion.

4.4 Project Limitations and Future Steps

Our dataset and machine learning (ML) model have limitations and could be improved in many ways as delineated in Figure 9. Due to our high precision and low recall, we could aim to reduce bias in our dataset by using methods such as k-fold cross-validation or increasing the number of layers in the CNN model.

One limitation of using sigmoid and binary cross entropy with logits is its inability to distinguish between examples that are strongly correlated or uncorrelated with the predicted class label. Loss functions such as these must employ binary cross entropy in conjunction with last-layer sigmoid, which occasionally results in numerical imprecision or instability [5]. It would be beneficial to explore additional methods to modify the loss function implementation and examine other techniques, such as Focal Loss, which directs more attention to incorrectly classified samples rather than easily classified samples [6]. As the model confidence increases, the loss would be reduced, and the model performance should increase as greater time would be directed to ensuring better classification with more challenging examples. Focal Loss would, more importantly, address the issue of class imbalance, especially if the majority class heavily dominates the loss and gradient descent process [7].

Furthermore, we aim to reduce our learning rate during validation, allowing more efficient training for the model. Implementing data augmentation, such as rotations or flips, and reducing the number of identified classes can also diversify the dataset and assist in creating a more robust model. Additionally, exploring the accumulation of a larger sample size for each chest diagnosis, including rare abnormalities, would likely decrease the number of incorrect diagnoses. This effort could also complement adding extra layers to the CNN.

In the future, creating an API (Application Programming Interface) for the ML model would allow the methods of classification to be more accessible for researchers in the biomedical fields to compare classification data and results more seamlessly as they progress toward integrating artificial intelligence (AI) into the field of radiology.

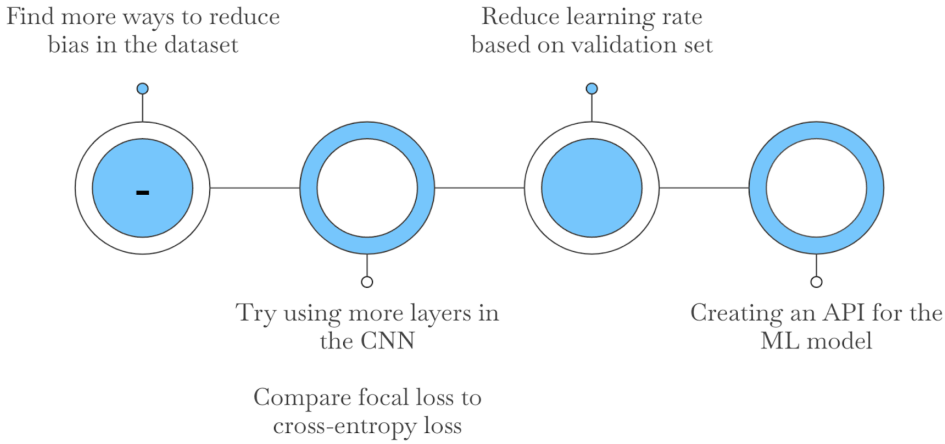

Figure 9: Flowchart delineating the future direction of the project. Process includes reducing dataset bias, adding additional layers to the CNN, attempting incorporation of different loss techniques (i.e. Focal Loss), reducing learning rate, and creating an API.

5 References

[1] Brady A. P. (2017). Error and discrepancy in radiology: inevitable or avoidable?. Insights into imaging, 8(1), 171–182. https://doi.org/10.1007/s13244-016-0534-1

[2] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., & Ni, B. (2023, January 19). MedMNIST v2 - a large-scale lightweight benchmark for 2D and 3D Biomedical Image Classification. Nature News.
https://www.nature.com/articles/s41597-022-01721-8

[3] Lozano, N. (2023, February 10). Can't get in to see the doctor? why there's a physician shortage and what's being done about it. KSBY News. https://www.ksby.com/news/local-news/cant-get-in-to-see-the-doctor-why-theres-a-physician-shortage-and-whats-being-done-about-it

[4] Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G., Menze, B., Ronneberger, O., Summers, R. M., Bilic, P., Christ, P. F., Do, R. K. G., Gollub, M., Golia-Pernicka, J., Heckers, S. H., Jarnagin, W. R., Cardoso, M. J. (2019, February 25). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv.org.
https://doi.org/10.48550/arXiv.1902.09063

[5] Hentschke, H. (2019, February 21). Sigmoid activation and binary crossentropy-a less than perfect match?. Medium. https://towardsdatascience.com/sigmoid-activation-and-binary-crossentropy-a-less-than-perfect-match-b801e130e31

[6] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2018, February 7). Focal loss for dense object detection. arXiv.org. https://arxiv.org/abs/1708.02002

[7] Nayak, R. (2022, April 28). Focal loss : A Better Alternative for Cross-Entropy. Medium. https://towardsdatascience.com/focal-loss-a-better-alternative-for-cross-entropy-1d073d92d075

[8] Shelke, A., Inamdar, M., Shah, V., Tiwari, A., Hussain, A., Chafekar, T., & Mehendale, N. (2021). Chest X-ray classification using Deep Learning for automated COVID-19 screening. SN computer science.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8152712/